

THE UTILITY OF CLASSIFICATION SYSTEMS IN
ORTHOPAEDIC SURGERY

CENTRE FOR NEWFOUNDLAND STUDIES

**TOTAL OF 10 PAGES ONLY
MAY BE XEROXED**

(Without Author's Permission)

ANDREW JOHN FUREY



The Utility of Classification Systems in Orthopaedic Surgery

By

Andrew John Furey MD

A Thesis Submitted to the School of Graduate Studies in partial fulfillment of the
requirements for the degree of Masters of Science

Faculty of Medicine
Memorial University of Newfoundland
St. John's, Newfoundland and Labrador

2004



Library and
Archives Canada

Bibliothèque et
Archives Canada

Published Heritage
Branch

Direction du
Patrimoine de l'édition

395 Wellington Street
Ottawa ON K1A 0N4
Canada

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

ISBN: 0-612-99075-3

Our file Notre référence

ISBN: 0-612-99075-3

NOTICE:

The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protègent cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

TABLE OF CONTENTS

	<u>Page</u>
Title page	i
List of Tables	v
List of Figures	vii
Acknowledgements	viii
Dedication	ix
Chapter 1- Introduction	
1.1 Background.....	1
1.2 Fracture Classifications.....	3
1.2.1 Traditional Classifications.....	7
1.2.2 Comprehensive Classifications.....	13
1.3 Factors Affecting Classifications.....	18
1.4 Analysis of Agreement.....	20
1.4.1 Interobserver Reliability.....	22
1.4.2 Intraobserver Reliability.....	22
1.4.3 Kappa Statistic.....	22
1.5 Literature Review.....	26
1.6 Objectives.....	31

Chapter 2- Study 1: Calcaneal Fractures.....	32
2.1 Design	33
2.2 Ethical Considerations.....	33
2.3 Abstract.....	34
2.4 Introduction.....	35
2.5 Methods.....	37
2.6 Results.....	38
 Chapter 3- Study 2: Subtrochanteric Femur Fractures.....	 41
3.1 Design.....	42
3.2 Ethical Considerations.....	42
3.3 Abstract.....	43
3.4 Introduction.....	44
3.5 Methods.....	46
3.6 Results.....	47
 Chapter 4- Study 3: Spinal Stenosis.....	 52
4.1 Design.....	53
4.2 Ethical Considerations.....	55
4.3 Abstract.....	56
4.4 Introduction.....	57
4.5 Methods.....	60
4.6 Results.....	61

Chapter 5- Discussion/Conclusion.....	66
5.1 Results.....	66
5.1.1 Calcaneal Fractures.....	66
5.1.2 Subtrochanteric Femur Fractures.....	68
5.1.3 Spinal Stenosis.....	71
5.2 Statistics.....	76
5.2.1 Analysis of Agreement.....	84
5.3 Conclusions.....	87
 References.....	 89

Appendix A: Questionnaire for Calcaneal Fractures

Appendix B: Human Investigation Committee Approval for Calcaneal Fractures

Appendix C: Questionnaire for Subtrochanteric Fractures

Appendix D: Human Investigation Committee Approval for Subtrochanteric Fractures

Appendix E: Questionnaire for Spinal Stenosis

Appendix F: Human Investigation Committee Approval for Spinal Stenosis

LIST OF TABLES

		Page
Table 1	Revised Trauma Score.....	4
Table 2	Gustilo-Anderson Classification of open fractures.....	9
Table 3	Observer variability in the orthopedic literature	27
Table 4	Total number of CT Scans classed according to types based on Sanders classification	39
Table 5	Total number of CT Scans classed according to subtypes based on Sanders classification according.....	39
Table 6	Total number of radiographs classed according to types based on Russell-Taylor classification according to individual observer after the first set of observations.....	49
Table 7	Total number of radiographs classed according to types based on Russell-Taylor classification according to individual observer after the second set of observations...	49
Table 8	Total number of radiographs classed according to subtypes based on Russell- Taylor classification according to individual observer after the first set of observations...	50
Table 9	Total number of radiographs classed according to subtypes based on Russell-Taylor classification according to individual observer after the second set of observations.....	50
Table 10	Mean values for classification of stenosis at L4-5.....	63
Table 11	The mean measurements of AP diameter of L4-5 for spines classified as normal, as well as the ANOVA.....	63

LIST OF TABLES

Table 12	The mean measurements of AP diameter of L4-5 for spines classified as mild, as well as the ANOVA.....	64
Table 13	The mean measurements of AP diameter of L4-5 for spines classified as moderate, as well as the ANOVA.....	64
Table 14	The mean measurements of AP diameter of L4-5 for spines classified as severe, as well as the ANOVA.....	65

LIST OF FIGURES

	Page
Figure 1 Neer's classification of proximal humerus fractures.....	11
Figure 2 AO comprehensive classification of fractures.....	15
Figure 3 Sanders Classification of intraarticular fractures of the os calcis based on CT Scans.....	40
Figure 4 Russell-Taylor Classification of subtrochanteric fractures of the femur based on AP and lateral radiographs.....	51
Figure 5 Schematic representation of canal changes with degenerative stenosis.....	54

ACKNOWLEDGMENTS

The completion of this project would not have been possible without the assistance, guidance and understanding of several individuals.

To the Division of Orthopaedic Surgery for their encouragement. Without their participation this undertaking would not have been possible. In particular, Dr. Craig Stone for his patience in allowing me to pursue this project while attending to resident responsibilities.

To all of the participants for their time in reviewing the endless number of images without compensation and sacrificing their own free time. In particular, for those who participated in more than one of these exercises, namely Dr. Craig Stone and Dr. Frank Noftall.

To my father George, for assisting in proof reading subjects he knew nothing about.

To my supervisory committee; Dr. Daniel Squire, Dr. Mark Borgonkar, and Dr. John Harnett, for their commitment and the giving of their time and effort in the completion of this project.

To my supervisor, Dr. Harnett, whose selfless leadership was greatly appreciated.

Thank you I am forever in your debt.

DEDICATION

This thesis is dedicated to my wife Allison, for her patience and understanding in completing this project. And to my parents, my grandfather J.B. O'Keefe and my Aunt Sister Rosalita Furey who have instilled in me the desire to ask questions and have helped foster the pursuit of answers.

CHAPTER 1 : INTRODUCTION

1.1 Background

Mankind has always had a tendency to classify: plants, animals, weather patterns or personality characteristics. The scientific literature is replete with examples of such classifications. From the complex classifications of solar systems to the judging of good, bad or ugly, man has an underlying desire to classify and organize. Modern medicine in particular has relied heavily on classification schemes. Classifications allow physicians to combine common pathologies, diseases, patient characteristics, lab values, etc. into groups with common characteristics. These groupings provide physicians with a means of communicating with a common understanding; they give the ability to provide prognosis and direct treatment, as well as providing useful research tools.

Orthopedic surgery has an abundance of classification systems. The orthopedic surgeon is often faced with the decision of whether to treat certain pathologies with surgery or without surgery. As a result, historically, orthopaedic surgeons have attempted to develop classifications of both traumatic (e.g. fractures) and non-traumatic disorders (e.g. osteoarthritis). The purpose of any medical classification is to group conditions with common characteristics together, thereby enabling the clinician to direct treatment, follow results of the chosen treatment, advise patients regarding prognosis and help future orthopaedic surgeons direct treatment.

Publications dealing with orthopaedic management use classifications to help answer the questions: given the classification of this pathology, what is the best treatment that will provide the best prognosis (Sanders, 1997); if the provided treatment fails, was it

the result of a poor treatment or the result of a misclassification? Herein lies an inherent problem of classification systems in general and orthopaedics in particular.

However, despite the potential for error and lack of reliability, classifications represent an important tool in orthopaedic surgery. They are necessary for publication of clinical research as a measure of severity and a predictor of outcome (Swiontkowski, Sands, Agel, Diab, Schwappach, & Kreder, 1997).

Classifications quoted in the literature have been used extensively but have not been assessed in terms of their reproducibility or validity. In order for a classification to be valid it must be both accurate and reliable. In order to achieve reliability there must be an adequate degree of interobserver and intraobserver reliability. In order to determine the degree of accuracy one must determine the agreement between preoperative assessment of classification and the intraoperative findings. There are several factors which affect the extent to which these criteria can be met. These include factors such as: the quality of the radiographs, the ability of the observer to identify the correct pathology and binary decision making (Drischl, & Adams, 1997).

Overall, orthopaedic surgery has been witness to an explosion of classification systems: some good, some not so good, and some that remain unstudied. Consequently, the study of classifications in orthopedic surgery has become an important part of orthopaedics.

1.2 Fracture Classification

Trauma is an area where classifications are particularly important in orthopaedic surgery. Many classification systems exist for the assessment of the acutely traumatized patient. Some of these classifications allow for the combination of subjective and objective assessment to help direct treatment, such as the classification of shock in the assessment of blood loss. Other classifications are more appropriately designed for the appraisal of prognosis, such as the Glasgow Coma Scale (Teasdale & Jennett, 1974). Still others are available for the assessment of the local institution's standard of care when comparing to national and international trauma registries such as the Revised Trauma Score (Table 1) (Greenfield, Mulholland, Oldham, Zelenock, & Lillemoe, 2001).

More than any other area of orthopedic trauma, fractures lend themselves to classification systems. In the past fractures have been classified according to the extent of bony injury, soft tissue injury, mechanism of injury and anatomic relationships. Although there have been numerous classifications of fractures, perhaps the one classification which first became widely used was the description of ankle fractures by Percival Pott. This system was popularized before the advent of radiographs (Pott, 1966). Pott, in 1769, in his *Remarks on Fractures and Dislocations*, described the proposed mechanisms of ankle fractures, their treatments, and their respective outcomes (Pott, 1966). Although used frequently into the 20th century, debate arose as to which fractures should be considered "Pott's" fractures.

Table 1 : Revised Trauma Score (Greenfield, Mulholland, Oldham, Zelenock, & Lillemoe,2001).

GCS	Systolic Blood Pressure	Respiratory Rate	Coded Value
13-15	>89	10-29	4
9-12	76-89	>29	3
6-8	50-75	6-9	2
4-5	1-49	1-5	1
3	0	0	0

Subsequently there began an explosion in the orthopedic literature of eponymous classifications of fractures of all bones of the skeleton. Some of these have been based on anatomy; others have been based on function. As a result, one type of fracture may have many different classification systems. Colton noted there were twenty new classifications for olecranon fractures between 1960 and 1981 (Colton, 1991). This haphazard explosion of classifications led to confusion and disarray in the orthopaedic literature. Individual surgeons were unable to compare results of independent studies using different classification systems. Many of these classifications were born out of surgeons' personal experiences, and most were not validated.

Publications dealing with fracture management use classifications to answer the question: what is the best treatment that will provide the best prognosis? Then, if treatment fails, is it the failure of the chosen management or is it the failure to provide the appropriate classification of the fracture? In fact, it has been suggested that fractures of a bone represent a continuum of pathologies, whereas drawing the categorical lines is arbitrary (Schwiontkowski, et al., 1997).

In order for a classification to be functional it must meet a number of conditions: it must accurately describe the nature of the injury and guide the treating surgeon; it must also establish a predictive role in terms of prognosis and outcome; and independent users should be able to discuss treatment and prognosis of individual injuries with a common understanding of the injury based on the classification (Garbuz, Bassam, Masri, Esdalie & Duncan, 2002).

Also, in order for a fracture classification to be valid it must meet certain criteria: it should be reliable and accurate; it should be reproducible between users such that a communication between users is possible without discrepancy (Garbuz, et al., 2002). Satisfying these criteria allows surgeons to conclude that the system is reliable and enables them to compare results of different treatments, and their prognoses. The validity of a classification refers to how accurately a classification captures the nature of the fracture; that is, how accurately it describes the true pathology (Wright & Feinstein, 1992). This essentially requires a judgment of preoperative radiographs and the intraoperative findings. This would require independent intraoperative assessment and represents an obvious bias and an underlying problem in validating classifications based on radiographs. As for any diagnostic test, validation of a classification must have a high degree of reliability (Garbuz et al., 2002).

The fracture classification system must be easy to interpret, since one of the purposes of classifying is to simplify. Some classifications have become so extensive that they lose their functional applicability.

Attempting to eliminate these problems, the orthopedic literature became less flooded with new classifications. Systems themselves were scrutinized for validity and reliability. Research has been directed towards the reproducibility of classifications between a single user at different times, (i.e. intraobserver reliability) and between different users, (i.e. interobserver reliability). However, several classification systems remain in constant use despite the fact they have not been validated.

Classification systems of fractures in orthopaedics can essentially be grouped into traditional classifications and comprehensive classifications.

1.2.1 Traditional Classifications

Many classifications of orthopaedic fractures involve the radiographic appearance of the fracture and a significant portion of the classifications available have not been validated. Nonetheless, these classifications continue to appear in the literature (Rockwood & Green, 1998). As more advanced classifications evolve, it is unlikely that the older classifications will disappear from the literature.

Traditional classifications may be descriptive with or without direct reference to treatment and prognosis. For example, many classifications use terminology such as 'name-grade-n' or 'name-type-x'. A good example of this is a Garden type 2 subcapital hip fracture (Garden, 1968).

Traditional classifications can be further divided on the basis of whether they are nominal, ordinal or scalar systems (Rockwood & Green, 1998). However, in reality, most classification systems involve a combination of these three systems.

Most of the traditional classifications are nominal in nature (Rockwood & Green, 1998). They use descriptive terminology to define common fracture lines. For example, a posterior wall hemi transverse -type fracture of the acetabulum conveys a common understanding that the fracture involves the posterior wall in addition to a transverse fracture of the anterior column of the acetabulum. Nominal classifications attempt to remain simple. They often are descriptive of the underlying anatomy, and how this anatomy has been disrupted. This, therefore, provides a surgeon with a guide as to how to correct the underlying anatomy.

Ordinal classifications are based on numeric classification of fractures usually based on the degree of severity of the injury, for example the Gustilo and Anderson classification of open fractures (Table 2). This represents a commonly used classification system in orthopaedic surgery. Gustilo-Anderson (1976) proposed that open fractures be reported as type I, type II or type III on the basis of the size of the wound, the extent of soft tissue injury and the degree of contamination. Type III was further subdivided into type III-A, III-B, and III-C according to the degree of soft tissue injury and the need for vascular reconstruction. Consequently, this is a true ordinal classification with type III worse than a type II, and a type III-C worse than a type III-A (Gustilo & Andersen, 1976). Despite the fact that this classification is regularly used to assess extent of injury, to determine the urgency of treatment, to determine the type of treatment provided, and to assess prognosis, it has been shown to be less than adequate. Brumback and Allan (1992) proved that this classification provided only moderate to poor interobserver agreement. They concluded that the classification is useful on an individual case basis but is less than adequate when comparing treatment modalities of different published series.

Scalar classifications involve measurements of fracture displacement. For example, in determining the stability of a cervical spine injury, the atlanto-dens interval (ADI) is often stated with specific measurements determining the degree of stability. In an otherwise healthy patient, an ADI of greater than 5mm implies instability which will require surgical stabilization whereas an

Table 2: Gustilo-Anderson Classification of open fractures

















Type	Wound	Mechanism	Contamination
Type I	< 1cm	Low energy	Minimal
Type II	1-10 cm	High energy	Moderate
Type III	>10cm	High energy	Extensive
Type IIIA	Minimal soft tissue stripping		
Type IIIB	Extensive soft tissue stripping		
Type IIIC	Vascular injury		

ADI of 3 to 5mm may indicate instability and an ADI of less than 3 mm indicates stability which can be treated without surgical stabilization (White, Panjabi, & Southwick, 1975).

Perhaps the most extensively studied traditional fracture classification is the Neer classification of proximal humeral fractures. The classification introduced by Neer (1970) outlined six classes of proximal humeral fractures based on the number of parts involved (Figure 1). The definition of a fracture fragment was greater than 1cm displacement or greater than 45 degrees angulation. Neer (1970) defined one part fractures as those which were minimally displaced and did not significantly disrupt the normal anatomy. He recommended that these fractures be treated non-operatively. Group II fractures were displaced fractures of the anatomic neck of the proximal humerus. Group III were defined as displaced fractures of the surgical neck. Group IV were defined as displaced fractures of the greater tuberosity. Group V were defined as displaced lesser tuberosity. Group VI was defined as a fracture dislocation.

Neer's work in classification of proximal humerus fractures arose from the previous existing classification systems and from the anatomy of the proximal humerus. Prior to Neer, attempts were made to classify these fractures based on the mechanism of injury and the underlying vascular anatomy (Dehne, 1945). Codman (1934) originally described a classification based on the vascular supply to the segments of the proximal humerus. Neer (1970) modified this classification by looking at the magnitude of injury involved. He hypothesized that the greater the degree of trauma the more displaced the

Figure 1: Neer's classification of proximal humerus fractures (1994).

Displaced Fractures				
	2-part	3-part	4-part	Articular Surface
Anatomical Neck				
Surgical Neck				
Greater Tuberosity				
Lesser Tuberosity				
Fracture-Dislocation	Anterior 	Anterior 	Anterior 	Anterior 
	Posterior 	Posterior 	Posterior 	Posterior 
Head-Splitting				

fracture would likely be. He defined a significant displacement as 1 cm and significant angulation as 45 degrees.

Neer's classification was originally widely accepted and utilized based on its utility as an anatomic classification which could direct treatment (Rockwood & Green, 2001). As a result of its popularity, it was the classification subject to the most scrutiny in orthopaedics. Original criticisms were centered around the prognostic indicator being the overall group and not necessarily the number of parts (Rockwood, & Green 2001). As a result, Neer refined his classification into a 4-part concept, that is, one-part, two-part and three-part fractures on the basis of the anatomy involved (Neer, 1994).

Many authors feel that it was largely Neer's classification which led orthopedic surgeons to objectively assess the utility of classification systems in the literature. In fact, several studies including Bernstein et al (1996), Brein et al. (1995), Kristiansen et al (1988), and Sidor et al (1993) have shown a lack of interobserver agreement in using Neer's classification of proximal humerus fractures. Neer recognized the scrutiny to which his classification was subject and in a letter to the editor of *Journal of Bone and Joint Surgery* suggested that the classification was not intended to be used as an exact system but to provide surgeons with conceptual groups when considering surgical planning and management (Neer, 1996).

Essentially, Neer's classification was one of the classifications which led orthopedic surgeons to look closer at the degree of reliability of classifications in the literature. It marked the end of the explosion of classifications of fractures that began in the middle of the century and forced investigators to critically review the classifications which were already present.

1.2.2 Comprehensive Classifications

The problems with the traditional classifications of fractures led to the desire to have a unified classification which could be applied to every bone. Professor Maurice Muller from the Association of the Study of Internal Fixation (ASIF) was a pioneer in developing such a classification system. He consulted with Arbeitsgemeinschaft fur Osteosynthesefragen (AO) and non AO fellows alike in pursuit of this system.

Muller himself said that “a classification is useful only if it considers the severity of the fracture and severity as a basis for treatment and for evaluation of the outcome of treatment” (Colton, 1991).

Muller, Nazarian, Koch and Schatzker (1990) decided the system needed to meet the following criteria:

- it must be logical and consistent;
- it must reflect the injury;
- it must be easy to recall;
- it must be comprehensible across countries and languages;
- it must be computer compatible.

Although such a system was desirable for simplifying the orthopedic literature, the development was met with much scepticism.

Despite this Muller et al. (1990) pressed on. He devised the following universal system. Long bones or pair of bones are assigned the following codes:

humerus	1
radius/ulna	2
femur	3
tibia/fibula	4

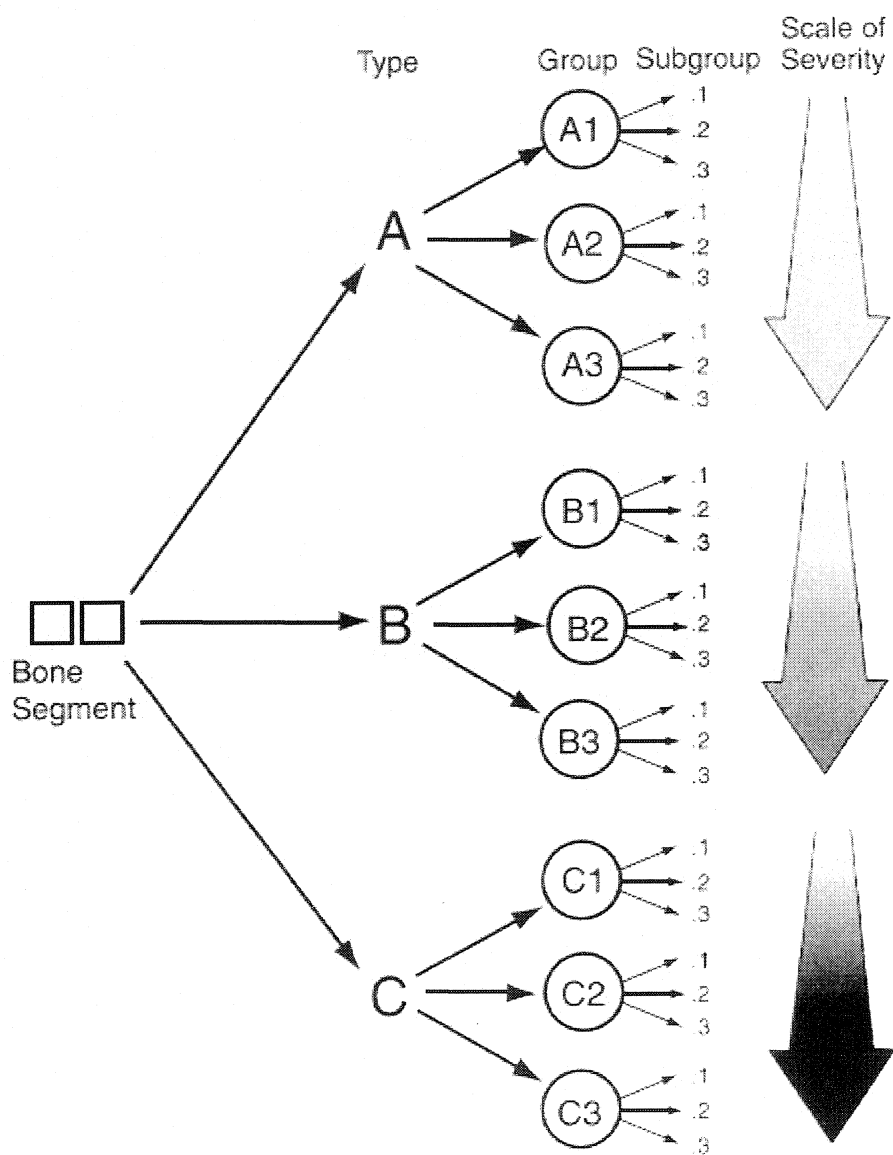
Next the level of each bone is assigned a code on the basis of one of four main zones:

proximal metaphyseal/articular	1
diaphyseal	2
distal metaphyseal/articular	3
malleoli	4

Following this scheme a mid-shaft radius fracture is considered a 2.2.

Furthermore diaphyseal fractures have qualifiers A, B, or C, based on the type of injury, for example, a type B represents a third butterfly fragment. Metaphyseal/articular fractures are divided into extra-articular A, unicondylar B, and bicondylar C. This classification has come to be known as Muller's Arbeitsgemeinschaft für Osteosynthesefragen (AO) Comprehensive Long Bone Classification System. Overall the most recent scheme of this classification can be summarized in Figure 2.

Figure 2: AO comprehensive classification of fractures (Muller et al, 1990)



The proposed advantage of this is that it will enable accurate comparisons of fracture treatments without the concerns of validity and reliability of the traditional classifications.

The Orthopedic Trauma Association (OTA) has developed a system modeled after the AO system (Anonymous, 1996). It is different in that it is developed as a dynamic classification which is subject to review and redesign every 3 years.

Despite the exhaustive attempts to design a universal classification which could be uniformly adapted throughout the orthopedic literature, there are still criticisms of these comprehensive classification systems. Some of the criticism is that this classification forces hierarchy on fractures which do not necessarily obey hierarchical organization. (Rockwood & Green, 1998) For example, proximal humerus fractures which are four parts are not differentiated from those which are three parts, despite differences in potential treatment modalities. Essentially some argue that it lacks the descriptive nature which is sometimes useful when considering surgical management. As well, some debate the reproducibility of this classification system as the more complicated and specific the attempt to numerically classify fractures the more subjectivity is implicated.

The reliability of the AO/OTA Fracture classification system has been assessed in terms of specific locations. For example, Swiontkowski et al.(1997) studied the interobserver reliability of AO classification of pilon fractures of the distal tibia. Their results showed a moderate degree of agreement. They concluded that researchers should continue to use and refine this classification. As well, Andersen, Blair, and Steyers (1996) analyzed interobserver variability of the AO classification of distal radius

fractures. Their results revealed only a fair degree of agreement. However, when they collapsed the system from 27 to 9 categories and further combined subclasses to 3 main categories, substantial agreement occurred (Andersen et al., 1996).

Overall, the comprehensive classifications attempt to reduce the descriptive nature of the traditional classification. However, they are plagued by problems similar to traditional classifications. Part of the problem is that surgeons are attempting to force a continuous variable with infinite possibilities, that is fracture patterns, into a dichotomous variable (Swiontkowski et al., 1997).

1.3 Factors Affecting Classification

The extent of observer variability is influenced by many factors. In order to improve the reliability one needs to identify and assess these factors so that they may be minimized for future classifications.

Garbuz et al. (2002) identified three potential areas of discrepancy in using a classification system. These included clinician variables, patient variables, and procedure or examination variables.

Clinician variation refers to variation between observers (Garbuz et al., 2002). This can result from variations in history taking, physical examination and radiographic interpretation. For example, in assessing scoliosis using Cobb's angles, the measurement depends on exactly where the physicians draw the lines and which vertebral bodies they choose. Inconsistency in selecting lines and bodies will lead to variations in measurements. In the case of Cobb's angle the variation has been shown to be 5 degrees for both interobserver and intraobserver variability (Dutton, Jones, Slinger, Scull, & O'Connor, 1989). Attempting to classify a continuous variable into an arbitrary category, one observer may take a borderline case to be one class while another observer may place the same borderline case in a different category.

Another factor impacting on classification systems is patient variation. This occurs when different patient factors leads to different interpretation of results. For example, a radiograph of a distal radius fracture in a severely osteoporotic patient may be classified differently than the same fracture in a non osteoporotic patient.

The last factor identified by Garbuz et al. (2002) was the examination variability. This refers to radiographic technique. For example different strength x-ray beams will provide different information regarding osteopenia. This factor can be reduced by adapting uniform examination techniques for an institution.

A study by Dirschl et al. (1997) asked participants to classify tibial plafond fractures in order to assess some of these factors. They asked the observers to assess the adequacy of the films provided. In fact there was no agreement on what was deemed an adequate film. Therefore, improving the radiographs may not improve reliability. They also asked observers to trace the fracture fragments on the radiographs prior to classifying them in an attempt to reduce clinician variability. However, this did not improve interobserver variability. Also in a separate study it was shown that training the observers did not improve reliability in assessing Neer's classification of proximal humeral fractures (Sidor, et al., 1993).

1.4 Analysis of Agreement

In order for a classification to be useful it must be simple, reliable and valid.

Reliability refers to the reproducibility within and between users. That is, there must be a agreement between users. Essentially, disagreement introduces a measurement error which may in turn influence the validity of a classification. Minimizing this measurement error will enhance the study's results.

Initially one must assess if the variable being assessed is a continuous or categorical variable. That is, is the measured variable based on an infinite arithmetic scale (Hulley & Cummings, 1988), as is the case with a continuous variable? Or, is it measured by classifying information into categories, as is the case with categorical variables?

Furthermore, if the measurement is a categorical variable one should assess if it is a nominal variable, one that does not assume order (e.g. color of a patients eyes), or, do the categories assume a rank order and, are therefore considered to be ordinal variables (e.g. severity of a patients pain as mild, moderate, or severe) (Hulley & Cummings, 1988).

While continuous variables offer the most information, certain measurements do not lend themselves to a continuous measurement, such as assessing fracture classifications.

Regardless of the decision on whether a variable is continuous or categorical, the classification can increase knowledge, reduce bias and provide a means for communication, provided they are done correctly.

Precision is achieved when a measurement is nearly the same each time it is measured. Precision therefore is closely related to the terms consistency and reliability.

There are three main sources of error in measurement precision (Hulley, & Cummings,

1988). These include the following: observer variability, subject variability, and instrument variability. For a continuous variable this can be assessed using statistics such as the standard deviation and standard error. For categorical variables statistics such as the kappa statistic are generally used. Hulley and Cummings (1988) recommended five approaches to help enhance the precision of a measurement: standardizing the measurement method, training the observer, refining the instruments, automating the instruments, and repetition.

Accuracy refers to the amount a measured variable represents that which it is supposed to represent. A highly accurate variable will be extremely useful in determining if a classification is valid. Accuracy is affected by bias in terms of observer bias, subject bias, and instrument bias (Hulley, & Cummings, 1988). Assessing accuracy is done by comparing the measurement of the variable to the gold standard. For example, comparing the preoperative classification of fracture to the intraoperative classification of a fracture. In addition to those proposed for improving accuracy, Hulley and Cummings (1988) suggested three means by which a measure could improve its accuracy. These include making an unobtrusive measure, blinding, and calibrating the instruments.

Therefore, in order to assess a classification system, one must assess the degree of precision and accuracy it provides. That is, one must assess the amount of agreement which exists between users to provide an assessment of reliability or precision. As well, in order to consider the amount of accuracy present one must investigate the amount of agreement between the preoperative and intraoperative diagnosis (i.e. between the measure and the gold standard).

Originally the efforts in statistical analysis of this agreement were based on percentage of agreement; that is, how often the observers ranked the same class for the same specimen. However, it was noted that this was a less than ideal situation to assess agreement (Cohen, 1960). Cohen argued that percent agreement did not account for agreement by chance alone. He proposed a statistic, kappa, which accounted for chance and eliminated it from the assessment of agreement.

1.4.1 Interobserver Reliability

Interobserver reliability refers to the amount of agreement between different users in using the same classification system. In classification of fractures this would refer to the degree of agreement of two or more independent surgeons assessing the same radiograph.

1.4.2 Intraobserver Reliability

Intraobserver reliability requires that the same observer arrive at the same conclusion with regards to a classification. This can be assessed through a number of similar means. A surgeon may assess numerous films, some of which are the same, and as a result, intraobserver variability can be assessed. Additionally a surgeon may be asked to classify the same set of films at different times.

1.4.3 Kappa Statistic

Initially, agreement was often assessed by using measurements such as percentage agreement. This value would assess the percentage of observations observers agreed upon. The Kappa statistic aimed at better determining the degree of agreement by

eliminating the degree of agreement due to chance alone (Cohen, 1960). Since its introduction, the kappa value has been used extensively in the orthopedic literature to assess agreement.

The role of chance is not insignificant when assessing the degree of agreement between observers. Whenever two outcomes are compared there is a chance that they will agree entirely by chance. Just as one has a 25% chance of guessing a multiple choice question, when there are four options, chance plays a role in assessing concordance.

Cohen (1960) developed a method to help eliminate the element of chance when assessing agreement. The index is the kappa statistic and it is the statistic of choice when considering nominal or dichotomous scales and it corrects for chance agreement. The formula is as follows:

$$K = \frac{po - pc}{1 - pc}$$

where po= observed proportion of agreement and pc is the probability of chance agreement calculated as shown. Therefore, the values of K range from 0, when there is no agreement to 1, when there is absolutely perfect agreement, and <0, if the agreement is less than one would expect by chance alone (Kramer, 1981). Kappa is normally used to measure concordance between two observers. If more than two exist it can also be used by doing pair wise comparisons.

The value of kappa achieved has a quantitative level of significance which Landis and Koch (1977) have suggested be the following:

Value of K	Strength of Agreement
<0	Poor
0-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.0	Almost perfect

Ordinal values imply order and therefore there is a degree of agreement even when there is disagreement. The statistic of choice for ordinal values is the weighted Kappa (Kw) (Kramer, 1981). This statistic is similar to Kappa but assigns weights on the basis of the extent of disagreement. The formula is as follows:

$$Kw = 1 - q_0/q_c$$

Note this formula uses q not p and therefore simplifies the formula using disagreement instead of agreement. The weights corresponding to various levels of disagreement are as follows: 0 is perfect agreement, 1 is disagreement by one category, 2 is disagreement by two categories etc (Kramer, 1981). Again it is the value of Kw that is achieved and not the p-value which is of more importance. In order for Kw to be quantitatively significant it should probably approach the magnitude of +0.5-0.6 (Kramer, 1981). Standard errors for weighted kappa can also be calculated.

There are also formulas which are available to estimate the sample size calculations for a weighted kappa analysis, the one used throughout the body of this paper is the simplest one. By increasing the sample size one may not necessarily change the result of the kappa analysis, however it will tighten the confidence interval surrounding it (Kramer, 1981).

1.5 Literature Review

As noted earlier, there are numerous studies of classifications in the orthopedic literature. Despite this, classifications are still used which have not been assessed regarding their reliability. Table 3 summarizes some of the exhaustive research on classifications present in the literature.

Many of these studies have made similar conclusions. One such conclusion is that for classifications to be useful in the treatment of orthopedic conditions, they must have an underlying degree of simplicity. However, most offer only a fair to moderate degree of observer reliability. Therefore, they serve as a framework when considering the individual treatment of a problem.

This has led some journals, such as *Journal of Orthopaedic Trauma* to conclude that, given the difficulties with classifications, authors should attempt to use the AO/OTA classification of fractures. For this journal in particular when using any other classification, authors must be able to show an agreement with a kappa of minimum of 0.55 (Sanders, 1997). They also suggest that any new classification which is proposed be validated prior to introduction in the orthopedic literature.

Despite these recommendations, new classifications continue to appear in the orthopedic literature without adequate validation, while traditionally used classifications have not been assessed for observer variability. This presents a continued problem in the orthopaedic literature. Table 3 represents only some of the studies with their respective kappa values. Obviously it is impossible to critically assess each and every one of these studies. However, it is clear that some trends exist. Most classifications fail to provide the

Table 3: Observer variability in the orthopedic literature

Author	System	Interobserver Kappa Statistic
Andersen GR	Older's distal radius	0.75
Andersen DJ	AO distal radius	0.252
Andersen DJ	Frykman distal radius	0.364
Andersen DJ	Melone distal radius	0.337
Andersen DJ	Mayo distal radius	0.428
Brady OH	Vancouver periprosthetic Fractures	0.60
Brumbeck RJ	Gustilo open fractures	0.60
Brien H	Neer proximal humerus fractures	0.45
Campbell DG	AAOS acetabular bone loss in revision surgery	0.11-0.28
Campbell DG	Gross acetabular bone loss in revision surgery	0.19-0.62
Campbell DG	Praposky acetabular bone loss in revision surgery	0.17-0.41
Chan PS	Schatzker tibial plateau	0.62
Craig WL	AO ankle	0.62-0.78
Cummings RJ	King scoliosis	0.44
Dirschl DR	Reudi tibial pilon	0.43
Haddad FS	AAOS femoral bone loss	0.12-0.29
Kreder HJ	AO distal radius	0.33
Lenke LG	King scoliosis	0.21-0.63
Martin JS	AO tibial pilon	0.38-0.60

Table 4 continued : Observer variability in the orthopedic literature

<u>Author</u>	<u>System</u>	<u>Interobserver Kappa Statistic</u>
Martin JS	Reudi tibial pilon	0.46
McCaskie AW	Quality of cement in hip arthroplasty	-0.04
Sidor ML	Neer proximal humerus fractures	0.43-0.58
Siebenrock KA	Neer proximal humerus fractures	0.25-0.51
Siebernock KA	AO proximal humerus fractures	0.26-0.49
Smith SW	Ficat: osteonecrosis	0.46
Swiontkoski MF	Ruedii tibia pilon fractures	0.46
Thomson NOB	Garden femoral neck fractures	0.39
Ward WT	Severin congenital hip dislocation	-0.01-0.42

adequate level of agreement of 0.55 suggested for the *Journal of Orthopaedic Trauma* (Sanders, 1997).

Three classifications which are used frequently in the orthopaedic literature include the classification of calcaneal fractures, classifications of the subtrochanteric femur fractures, and the classification of the degree of lumbar spinal stenosis. These classifications are often quoted to provide insight into treatment and prognosis of these orthopaedic conditions.

The literature was thoroughly reviewed, using combinations of Medline (1966-2001) and Cochrane Library, specifically for the classification of os calcis fractures proposed by Sanders. Although there were many references to the use of Sanders classification of calcaneal fractures, there was no review of its reliability.

The literature was also reviewed, using a combination of Medline (1966-2002) and Cochrane Library, specifically for the classification of subtrochanteric femur fractures proposed by Russell and Taylor. There was no study assessing the degree of observer reliability.

Finally, the literature was reviewed, using a combination of Medline (1966-2002) and the Cochrane Library, specifically searching for the assessment of lumbar spine stenosis using computerized axial tomography scans (CT scans) and the measurement of the space for the spinal canal in determining the degree of stenosis. Although there was reference to what was thought to be the normal canal diameter, and the stenotic canal diameter there was no review of these numbers in terms of interobserver variability.

In summary there are many studies assessing many different classifications of orthopedic pathologies. There were no studies assessing the degree of variability when

using Sanders classification of calcaneal fractures. There were no studies assessing the degree of variability when using the Russell-Taylor classification of subtrochanteric fractures. Finally, there were no studies assessing the degree of variability when measuring canal diameter and determining which lumbar spines were stenotic.

1.6 Objectives

For the purpose of this thesis it was decided to focus on the reliability of classifications in helping to assess the validity.

The lack of literature in the areas of calcaneal fractures, subtrochanteric fractures of the femur, and spinal canal stenosis lead to the primary goals of this study:

1. to determine the degree of interobserver in using Sanders classification of calcaneal fractures;
2. to determine the degree of interobserver and intraobserver reliability in using the Russell-Taylor classification of subtrochanteric fractures of the femur;
3. to determine the degree of interobserver and intraobserver reliability in measuring spinal canal diameter in assessing lumbar spine stenosis;
4. to determine the degree of interobserver and intraobserver reliability in classifying spinal canal stenosis of the lumbar canal as mild, moderate, or severe.

As previously addressed validity is difficult to assess for classification systems of fractures. It was the intention of these studies to focus on one element of validity, reliability. To further assess validity, accuracy of these classifications should be assessed. This perhaps should be the focus of future studies. This would allow then for both elements of validity, reliability and accuracy, to be assessed,.

Chapter 2 : Calcaneal Fractures

The os calcis, also known as the calcaneus, or the heel bone, is the main bone of the hind foot. It serves as a structure which must support the weight of the body. It must also provide for adequate motion through the subtalar joint to allow for locomotion over uneven ground. The calcaneus itself is a thin cortical shell of bone surrounding inner cancellous bone. The calcaneus can be divided into the tuberosity, or the posterior aspect, the body which is just anterior to the tuberosity, the lateral process, the anterior process, and the sustentaculum tali. The articular anatomy of the calcaneus is essential in understanding and treating fractures. The posterior facet is the largest articulating surface of the calcaneus. The posterior facet in combination with the anterior facet and the middle facet make up the subtalar joint. There is also the distal articular surface which is responsible for articulating with the cuboid bone anteriorly.

The calcaneus may be fractured in numerous ways including but not limited to falls from heights, brake pedal injuries, or twisting injuries. These fractures may be intraarticular (70-75%) or extraarticular (25-30%) (Kundell, Brutscher et al., 1964). Most classifications of the calcaneus have been based on the anatomy and are used to direct the treatment of fractures of the calcaneus. One of the most frequently used classifications of the calcaneus is the one proposed by Sanders (1992).

2.1 Design

The literature was reviewed and no studies assessing the reliability of Sanders classification of calcaneal fractures were found. The literature was reviewed regarding assessment of other classifications of fractures in the orthopedic literature. These methods were used as a guideline for assessing the Sanders classification. The design, which is reflected elsewhere in the literature, asked independent surgeons to classify a number of different radiographs based on Sanders' classification using a questionnaire (Appendix A).

2.2 Ethical Considerations

Permission to pursue this research was granted by the Human Investigation Committee of the Faculty of Medicine, Memorial University of Newfoundland after the appropriate review (Appendix B).

The underlying ethical issues which were considered when constructing the methodology of this project were based on the protection of the patient's identities. The cases of calcaneal fractures were reviewed from the Health Care Corporation of St. John's over a five year period. CT scans were selected and were coded such that the reviewers were unable to identify the patient from the radiographs.

Consent was obtained from each reviewer prior to reviewing the CT scans after the purpose of the research had been explained to the participants.

2.3 Abstract

The os calcis is the most frequently fractured tarsal bone. In 1992 Sanders developed a classification system based on coronal and axial CT scans of the calcaneus (Sanders, 1992). This classification is the one used most frequently today in treatment decision making and reporting of results. The objective of this study was to assess the degree of inter-observer variability in using this classification system. Thirty CT scans of calcaneal fractures were randomly chosen from the past five years in two tertiary care centers. The CT scans were reviewed by three orthopaedic surgeons and one senior orthopaedic resident who classified the fractures according to Sanders' classification. The results were first tabulated and analyzed using a weighted kappa test including the subcategories. The weighted kappa achieved was 0.56 with a 95% confidence interval of (0.45, 0.67). The subcategories of the classification were then further combined and a second weighted kappa was performed to assess agreement between general classes. The weighted kappa achieved was 0.48 with a 95% confidence interval of (0.37, 0.59). It was concluded that Sanders' classification system did achieve moderate agreement among users, thus representing a useful classification system.

2.4 Introduction

The os calcis is the most frequently fractured tarsal bone (Rockwood&Green, 1998). The complications resulting from such a fracture are numerous and include malunion, post-traumatic subtalar arthritis, chronic foot pain, peroneal tendonitis, and lateral impingement syndrome (Myerson & Quill, 1993). There have been many attempts to accurately describe and classify fractures of the os calcis. Classifications range from using the mechanism of injury, as Bohler did in the 1930's, to Essex-Lopresti's analysis of location of the fracture (Sanders, Fortin, DiPasquale, & Walling , 1993) (Giachino, Uhthoff, 1989). Despite exhaustive methods of analyzing and classifying these fracture patterns there were often less than ideal outcomes for patients with intraarticular fractures of the os calcis. Surgeons were often in disagreement as to which classification to use as well as when and which approach to use in operative intervention of these fractures (Sanders, et al., 1993). In an attempt to reduce confusion and clarify classifications of these fractures, Sanders (1992) developed a classification system of intraarticular fractures based on coronal and axial CT Scans of the fracture (Sanders, et al., 1993). This classification was developed in part to help stratify patients and thereby assist in deciding which patients required surgical intervention and which patients could be treated conservatively. Sanders used his classification to help predict the prognosis of these fractures and found that those patients who were best treated operatively were those who had type II and type III fracture patterns (Sanders, 1992). Given that Sanders classification is frequently used by surgeons and the classification may indeed affect

potential therapeutic decisions, it is essential that the classification system be applied uniformly and consistently. Hence the purpose of this study is to assess the degree of inter-observer variability in using Sanders classification system.

Sanders' classification is based on coronal and axial CT scans of the calcaneus (Fig. 3). Sanders described using the CT cut which displayed the undersurface of the posterior facet of the talus, at its widest point (Sanders, et al. 1993). Subsequently the posterior facet is divided into three sagittal columns of equal size by two lines A and B. This results in the posterior facet being divided into three columns, namely: medial, lateral and central (Sanders, et al., 1993). A third line C is also considered. It runs along the medial edge of the posterior facet and represents a line separating the sustentaculum tali from the medial edge of the posterior facet (Sanders, et al., 1993). Consequently three potential primary fracture lines are created, beginning laterally with A and moving medially to C, thereby creating four potential fracture fragments overall.

The classification is divided into four types. Type I represents all undisplaced fractures, regardless of the number of pieces involved. Type II represents two part fractures of the posterior facet and can be divided into three subtypes based on the primary fracture line, these include IIA, IIB, and IIC (Fig. 3)(Sanders, et al., 1993). Type III represents three part fractures with a centrally depressed fragment. Again this class can be subdivided into three subclasses based on the fracture lines, IIIAB, IIIAC, and IIIBC (Fig. 3). Type IV fractures represent highly comminuted fracture patterns.

2.5 Methods

Fifty CT scans of calcaneal fractures were reviewed from the databases between 1995 and 2000. Twenty-nine patients representing thirty intra-articular calcaneal fractures were selected on the basis that they were available and represented fractures that were classifiable according to Sanders classification. The CT scans were selected by the author who was not a reviewer and thus independent of the results. Although the CT scans were selected from the institution where the reviewing surgeons practice, the reviewers were blinded as to patients' names, treating surgeons, and the treatments the patients received.

The CT Scans, including the entire sheet of slices, were distributed to three orthopedic surgeons from the department of Orthopaedic Surgery at our institution, as well as one senior orthopaedic resident. The participants were also provided with a figure describing Sanders' classification and asked to classify each fracture based on Sanders' classification.

The results were subsequently tabulated excluding one of the CT's after it was realized that it was in fact a sustentaculum fracture and not classifiable with Sanders' classification. The classes and subclasses were treated as ordinal values and consequently a weighted kappa was chosen as the analysis of choice. Using the computer program PC Agree two separate weighted kappa tests were performed using first the entire class and then the subclasses. Ninety-five percent confidence intervals were calculated for both weighted kappa values.

A weighted kappa is a test of partial agreement between observers. It is most useful in categorical data. Designed originally by Cohen, the amount of agreement is assigned a specific weight and subsequently a kappa is calculated. Values range from -1 to $+1$ with a value of 0 representing the agreement by chance alone (Kramer, 1981). Assuming competent observers, a weighted kappa approaching $+0.5$ - 0.6 represents an acceptable degree of agreement (Kramer, 1981).

2.6 Results

The four surgeons' results were tabulated in terms of both classes and subclasses and the results are tabulated in Table 4 and Table 5. Using the computer program PC Agree the weighted kappa obtained using the classes as a whole was 0.48 , with a standard error of 0.058 and a subsequent 95% confidence interval of 0.37 to 0.59 . The weighted kappa obtained maintaining the subtypes was $k = 0.56$ with a standard error of 0.058 and subsequent confidence interval of 0.45 to 0.67 .

Table 4 : Total number of CT Scans classed according to types based on Sanders classification according to individual observer.

OBSERVER	TYPE I	TYPE II	TYPE III	TYPE IV
1	1 (3.45%)	12 (41.38%)	9 (31.03%)	7 (24.18%)
2	5 (17.24%)	15 (51.72%)	4 (13.79%)	5 (17.24%)
3	1 (3.45%)	23 (79.31%)	5 (17.24%)	0 (0%)
4	0 (0%)	15 (51.72%)	5 (17.24%)	9 (31.03%)

Table 5: Total number of CT Scans classed according to subtypes based on Sanders classification according to individual observer.

Observer	TYPE I	TYPE IIA	TYPE IIB	TYPE IIC	TYPE IIIAB	TYPE IIIAC	TYPE IIIBC	TYPE IV
1	1	10	2	0	9	0	0	7
2	5	12	3	0	2	1	1	5
3	1	14	9	0	4	1	0	0
4	0	10	4	1	4	0	1	9

Figure 3 : Sanders Classification of intraarticular fractures of the os calcis based on CT Scans.



Chapter 3 : Subtrochanteric Fractures of the Femur

The subtrochanteric portion of the proximal femur is an area which is often grouped with other fractures of the proximal femur as hip fractures. In fact, the subtrochanteric fracture behaves as an intermediate between a hip fracture and a femur fracture. The subtrochanteric region is defined as the area of bone distal to the intertrochanteric line, that is between the lesser trochanter and the isthmus of the femur. This region of the femur is subject to deforming forces which contribute to the difficulty in treating these fractures. The proximal portion is influenced by the pull of the abductors, short external rotators and flexors, while the distal end is pulled by the adductors. This presents a challenge to surgical reduction and internal fixation. These fractures have been the subject of much controversy in the orthopedic literature in terms of their treatment. This ranged from original description of internal fixation by Hoglund (1917) and Groves (1918) to more sophisticated options including intramedullary nailing such as described by Zickel (1967). Sanders and Regazzoni (1989) reported on the use of 95 degree condylar blade plate.

In attempts to direct the surgeon towards the correct treatment several classifications of subtrochanteric fractures were developed. Originally Boyd and Griffin (1949) described a classification that was followed by the widely used Seinsheimer (1978) classification. This however was proven to be inadequate (Gehrhen, Neilsen, Olesen et al., 1997). Subsequently, Russell-Taylor (1998) developed a classification. This classification is currently widely used, but has not been assessed regarding reliability.

3.1 Design

The literature was reviewed and no studies assessing the reliability of Russell-Taylor classification of subtrochanteric fractures were found. The literature was reviewed regarding assessment of other classifications of fractures in the orthopedic literature, these methods were used as a guideline for assessing the Russell-Taylor classification. The design which is reflected elsewhere in the literature asked independent surgeons to classify a number of different radiographs based on the Russell-Taylor classification using a questionnaire (Appendix C).

3.2 Ethical Considerations

Permission to pursue this research was granted by the Human Investigation Committee of the Faculty of Medicine, Memorial University of Newfoundland after the appropriate review (Appendix D).

The underlying ethical issues which were considered when constructing the methodology of this project were based on the protection of the patient's identities. The cases of subtrochanteric fractures were reviewed from the Health Care Corporation of St. John's over a period from 1995-2002. X-rays were selected and coded such that the reviewers were unable to identify the patient from the radiographs.

Consent was obtained from each reviewer prior to reviewing the x-rays after the purpose of the research had been explained to the participants.

3.3 Abstract

Subtrochanteric femur fractures often represent a therapeutic dilemma. With the advent of intramedullary nails in the 1980's, Russell and Taylor (1998) developed a classification system based on extension of the fracture to the piriformis fossa and lesser trochanter. This classification is frequently used today in treatment decisions, especially when one is considering use of an intramedullary nail. The objective of this study was to assess the degree of interobserver and intraobserver variability in using this classification system. Sixteen plain radiographs of subtrochanteric femur fractures were randomly chosen from the past five years in two tertiary care centers. The radiographs were reviewed by three orthopedic surgeons and one senior orthopedic resident who classified the fractures according to the Russell-Taylor classification. The data were collected and after a period of time the observers were asked once again to classify the fractures. The results were first tabulated and analyzed using a weighted kappa test including the subclasses. The weighted kappa achieved was 0.31 with a 95% confidence interval of (0.15,0.47). The subclasses of the classification were then further combined and a second weighted kappa was performed to assess agreement between categories as a whole. The weighted kappa achieved was 0.056 with a 95% confidence interval of (-0.089,0.20). The results of the second set of results was tabulated and a similar analysis was performed. The weighted kappa not including subclasses was 0.12 and, when considering the subclasses was 0.32. A final analysis was performed to determine the degree of

intraobserver variability. It was concluded that the Russell-Taylor classification system proved to achieve only minimal agreement among users and between users.

3.4 Introduction

Fractures around the hip represent a significant source of morbidity. The types of hip fractures can be identified as follows: femoral neck fractures, intertrochanteric fractures, and subtrochanteric fractures. The subtrochanteric fractures can occur in the elderly in a low energy injury or in a younger population through a high energy injury. These fractures often represent treatment dilemmas with options ranging from intramedullary nails to open reduction internal fixation with screws and plates. In 1978 Seinsheimer classified subtrochanteric fractures (Rockwood & Green, 1998) (Seinsheimer, 1978). This classification system was proven to be unreliable between users (Gehrhen, Neilsen, Olesen et al., 1997). A new classification developed by Russell and Taylor emerged in the mid 1980's. This classification is thought to be one of the most commonly used by Orthopedic surgeons who advocate treating these fractures with intramedullary nails. However this classification has not been shown to be valid in terms of interobserver or intraobserver variability. The purpose of this study was to assess the degree of interobserver and intraobserver variability in using the Russell-Taylor classification of subtrochanteric fractures.

This classification was developed in part to help stratify patients and thereby assist in deciding which patients would benefit from a femoral nail. The classification is simple in its concept and has two main classes with two subclasses per class. The essential determinant, of the classification and therefore potential treatment options, is the extension of the fracture line into the piriformis fossa. It disregards the degree of comminution and emphasis is placed on continuity of the lesser trochanter and extension of the fracture into the piriformis fossa (Russell & Taylor, 1998). If the piriformis fossa is not involved, regardless of the comminution, an intramedullary device can be used safely (Russell & Taylor, 1998).

Essentially the classification divides the fractures into those which do not involve the piriformis fossa, type I, and those which involve the piriformis fossa, type II. Type I is further subdivided based on involvement of the lesser trochanter. Therefore, type IA is a fracture which does not extend to the piriformis fossa and is located entirely below the lesser trochanter, whereas type Ib does not involve the piriformis fossa but does involve the lesser trochanter. Type II fractures extend to the piriformis fossa. Similarly, type IIa does not involve the lesser trochanter whereas type II b does (Russell & Taylor, 1998)

Figure 3.

Debate still exists as to which form of surgical intervention is the best treatment for these complicated fractures. This classification attempts to help clarify some issues surrounding this fracture pattern.

Given that the Russell-Taylor classification is frequently used by surgeons and the classification may indeed affect potential therapeutic decisions, it is essential that the classification system be uniform and consistent between surgeons. Hence, the purpose of

this study is to assess the degree of interobserver and intraobserver variability in using the Russell-Taylor classification system.

3.5 Methods

Prior to the start of the study the number of radiographs required was calculated to be 16 in accordance with rules for kappa analysis (Cicchetti, 1977). Cicchetti (1977), proposed that one could estimate the number of cases needed to determine a kappa by squaring the number of categories and multiplying by four. Subsequently radiographs of subtrochanteric femoral fractures were reviewed from the databases between 1995 and 2000. The radiographs were reviewed and sixteen radiographs with AP and lateral views were selected on the basis that they were thought to represent the normal population of the fracture patterns. The reviewers were blinded as to patients' names and treatments.

The radiographs were distributed to three orthopaedic surgeons from the Division of Orthopaedic Surgery at our institution, as well as one senior orthopaedic resident. The participants were also provided with a figure describing the Russell-Taylor classification and asked to complete a questionnaire asking them to classify each fracture based on the Russell-Taylor classification. An average of two months later the same radiographs were redistributed to the reviewers and they were again asked to classify the fractures.

The results were subsequently tabulated. A weighted kappa analysis was performed to determine the degree of interobserver variability on both occasions. A kappa analysis was used to assess intraobserver variability.

3.6 Results

For both sets of observations the surgeons' results were tabulated in terms of both classes and subclasses and are summarized in Table 6, Table 7, Table 8, and Table 9.

Treating the classes and subclasses as ordinal values a weighted kappa was chosen as the analysis of choice to determine interobserver variability. Two separate weighted kappa tests were performed considering the classes alone and then the subclasses. This analysis was performed for both sets of observations.

The weighted kappa obtained using the classes as a whole for the first set of observations was 0.056, and a subsequent confidence interval of (-0.089 to 0.20). The weighted kappa obtained using the classes as a whole for the second set of observations two months later was 0.12, and a subsequent confidence interval of (-0.039 to 0.29).

The weighted kappa obtained maintaining the subtypes for the first set of observations was 0.31 with a standard error of 0.080 and subsequent confidence interval of (0.15 to 0.47). The weighted kappa obtained maintaining the subtypes for the second set of observations was 0.32 with a standard error of 0.086 and subsequent confidence interval of (0.15 to 0.49).

The results were then again tabulated and an analysis of intraobserver variability was performed using a kappa analysis when considering both classes as a whole and considering the subclasses.

When considering the classes as a whole the following results were obtained. Observer one achieved a kappa of 0.73 with a standard error of 0.18. Observer two achieved a kappa of 0.35 with a standard error of 0.24. Observer three achieved a kappa

of 0.88 with a standard error of 0.000. Observer four achieved a kappa of 0.20 with a standard error of 0.18.

When considering the subclasses the following results were obtained. Observer one achieved a kappa of 0.64 with a standard error of 0.15. Observer two achieved a kappa of 0.28 with a standard error of 0.17. Observer three achieved a kappa of 0.77 with a standard error of 0.14. Observer four achieved a kappa of 0.42 with a standard error of 0.16.

Table 6 : Total number of radiographs classed according to types based on Russell-Taylor classification according to individual observer after the first set of observations.

Observer	TYPE I	TYPE II
1	10	6
2	10	6
3	16	0
4	10	6

Table 7: Total number of radiographs classed according to types based on Russell-Taylor classification according to individual observer after the second set of observations.

Observer	TYPE I	TYPE II
1	10	6
2	9	7
3	14	2
4	15	1

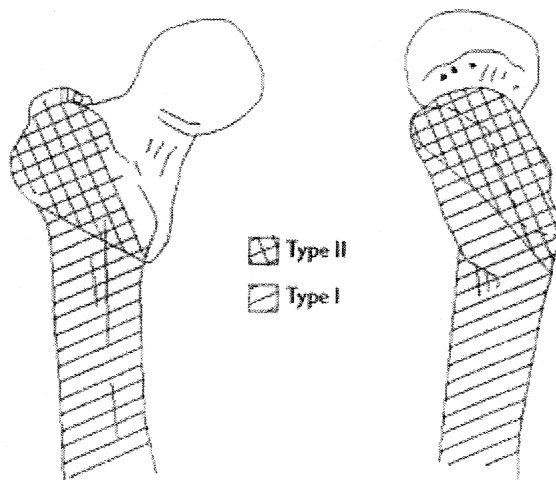
Table 8: Total number of radiographs classed according to subtypes based on Russell-Taylor classification according to individual observer after the first set of observations.

OBSERVER	TYPE IA	TYPE IB	TYPE IIA	TYPE IIB
1	3	7	2	4
2	4	6	1	5
3	6	10	0	0
4	3	7	0	6

Table 9: Total number of radiographs classed according to subtypes based on Russell-Taylor classification according to individual observer after the second set of observations.

OBSERVER	TYPE IA	TYPE IB	TYPE IIA	TYPE IIB
1	4	6	1	5
2	6	3	0	7
3	6	8	0	2
4	5	10	0	1

Figure 4: Russell-Taylor Classification of subtrochanteric fractures of the femur based on AP and lateral radiographs.



Chapter 4 Spinal Stenosis

Degenerative lumbar spinal stenosis was originally described in 1803 by Antoine Portal in hunchbacks with rickets as “too narrow vertebral canals”. In the 1950’s the concept of an acquired narrowing of the canal through a degenerative process was hypothesized by Verbiest (1975) (Hilbrand, 1999). The incidence of this often crippling disease has been reported from 1.7% to 8% of people over the age of fifty (Robertson, Llewellyn, & Taveras, 1973) (De Vilier, & Booyesen, 1976).

The pathophysiology of the disease follows the process of a degenerative pattern. It begins with normal changes in the spinal elements as we age. The disc becomes dehydrated and worn resulting in a loss of disc height. Subsequently the ligaments including the ligamentum flavum retain their length but effectively shorten and buckle resulting in a crowding of the spinal canal. These changes then affect the posterior elements placing abnormal stress across the facet joints resulting in arthritic changes and subsequent hypertrophy of the facets which further crowds the canal space. As a consequence there is a reduction in the canal diameter (Figure 5).

Patients will often report symptoms of back pain, claudication, leg pain, weakness, and voiding difficulties (Amundsen, Weber, Lilleas, Nordal, Abdelnoor, & Magneas, 1995). The most common radiculopathy patterns include L5 in 91%, S1 in 63%, L1-L4 in 23% and S2-S5 in 5% (Amundsen, et al., 1995).

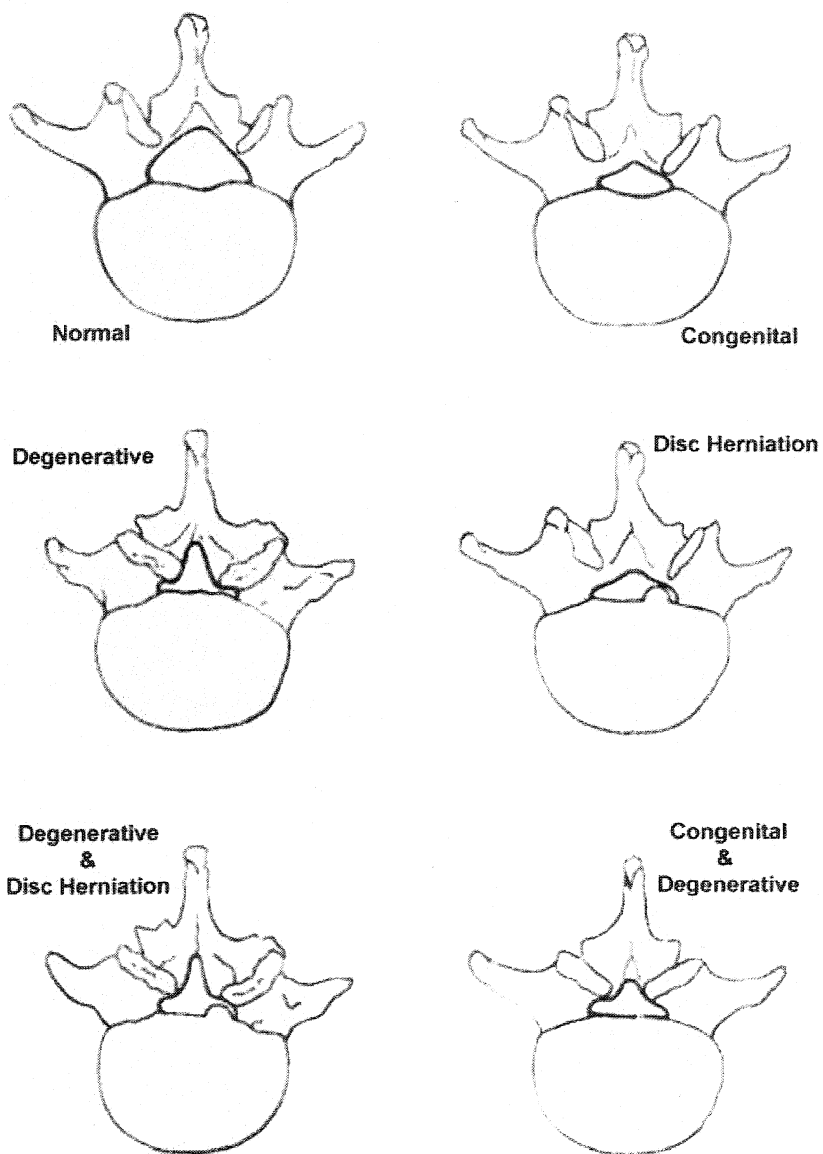
Diagnostic modalities used in assessing such patients include the following: plain film radiographs, myelography, CT scan, postmyelographic CT, and MRI. There are

several radiographic parameters which may be measured in assessing the spine for stenosis. These include assessment of the bone canal dimensions, as well as the dimensions of the cord itself. One of the problems associated with diagnostic imaging of the spinal canal is that there are a large portion of asymptomatic individuals who display evidence of lumbar stenosis. Boden et al. (1990) found 21% of individuals between the ages of 60 and 80 had MRI evidence of stenosis of the lumbar spine without clinical symptoms. It has been shown that the degree of stenosis evident on postmyelographic CT and MRI correlates with the intraoperative assessment of stenosis (Modic, Masaryk, Boumpfrey, Goormastic, & Bell, 1986). There has not, however, to knowledge of the author been any assessment of the degree of variability in measuring parameters for stenosis.

4.1 Design

The literature was reviewed and no studies assessing the reliability measuring the spinal canal were found. The literature was reviewed regarding assessment of other measurements and classifications of diseases in the orthopaedic literature. These methods were used as a guideline for assessing measurements. Independent reviewers were used to measure and classify a number of different CT scans based on the anterior-posterior, and interpedicular distances and to classify lumbar spinal stenosis using a questionnaire (Appendix E).

Figure 5: Schematic representation of canal changes with degenerative stenosis.



4.2 Ethical Considerations

Permission to pursue this research was granted by the Human Investigation Committee of the Faculty of Medicine, Memorial University of Newfoundland after the appropriate review (Appendix F).

The underlying ethical issues which were considered when constructing the methodology of this project were based on the protection of the patient's identities. The cases of spinal stenosis were reviewed from the Health Care Corporation of St. John's over a five year period. Patients with CT scans were randomly selected and were coded such that the reviewers were unable to identify the patient from the radiographs.

Consent was obtained from each reviewer prior to reviewing the CT scans after the purpose of the research had been explained to the participants.

4.3 Abstract

Diagnosis of lumbar spinal stenosis is a common clinical problem, as it represents both an anatomic and clinical diagnosis. Although the gold standard is considered to be myelography, many new tests have developed which are less invasive and offer potentially more preoperative information. The purpose of this study was to assess the degree of reliability in using one of these modalities, namely CT scanning. Twenty five CT scans reported as being stenotic were randomly chosen from the past five years in two tertiary care centers. The CT's were reviewed by two orthopedic surgeons, one senior orthopedic resident, two radiologists and one neurosurgeon who measured the AP and IP diameters of marked images representing L3-4, L4-5, and L5-S1. They were also asked to classify the spinal canal as normal, mild, moderate or severe stenosis. The results were first tabulated and analyzed using a weighted kappa test for the spinal canal as a whole. The weighted kappa achieved was 0.51 with a 95% confidence interval of (0.43,0.60). A second weighted kappa was performed assessing the agreement between orthopaedic surgeons (0.58) and between radiologists (0.58). Weighted kappas were performed considering the Verbiest classification at all three levels L3-4, L4-5, and L5-S1 were 0.25, 0.22, and 0.26 respectively. A series of four ANOVA analyses were performed to determine the degree of agreement in the AP measurements when considering how the scans were originally classified as normal, mild, moderate or severe. It was concluded that the degree of agreement between observers in assessing the degree of stenosis for the spine overall was moderate while the agreement between measurements and assessing individual level was less reliable.

4.4 Introduction

Degenerative lumbar spinal stenosis is a debilitating disease for many patients. Incidence increases with age, with a reported incidence between 1.7% and 8% in patients over 50 years of age (Robertson, Llewellyn, & Taveras, 1973) (De Vilier, & Booyesen, 1976). Stenosis was originally described in 1803 by Antoine Portal in hunchbacks with rickets as “too narrow vertebral canals”. In the 1950’s the concept of an acquired narrowing of the canal through a degenerative process was hypothesized by Verbiest (1975) (Hilbrand, 1999). The disease is characterized by degeneration of the spinal elements starting with the degeneration of the disc, resulting in loss of disc height, infolding of the ligamentum flavum, bulging of the annulus and arthritic changes within the facets. The consequence is a narrow canal and compression of the neural elements either centrally, or laterally with a narrowed foramen.

Degenerative lumbar spinal stenosis is both a clinical and anatomical disease which often presents a diagnostic dilemma. Several attempts have been made to arrive at one single radiographic test which would confirm the diagnosis. However, given the clinical nature of the disease and the incidence of stenotic spinal anatomy in asymptomatic individuals, no single test has proven to be the best. Myelography however, is considered to be the gold standard (McCulloch, & Transfeldt, 1997).

Patients will often report symptoms of back pain, claudication, leg pain, weakness, and voiding difficulties (Amundsen, 1995). Treatment options include nonoperative modalities such as non-steroidal anti-inflammatories, physical therapy, bracing, epidural steroid injections, facet joint injections and manipulations. Surgical options include lumbar decompression alone or in combination with a fusion with or

without instrumentation. The Maine Lumbar Spine Study (Atlas, & Keller, 1996) suggested that surgical intervention in patients who have failed conservative management of their lumbar spinal stenosis offers good short and intermediate term results. However they warned the long term results required further evaluation.

In order to treat stenosis correctly, either conservatively or with surgical intervention it must be correctly diagnosed. Several imaging modalities have been used to help diagnose and direct the treatment of lumbar spine stenosis.

There has been much controversy over establishing adequate, quantitative criteria for diagnosing the disease. Much of this has been focused on the modalities of CT scans and MRI (McCulloch, & Transfeldt, 1997). Several parameters have been proposed in assessing stenosis of the canal. They include measurements of the anterior-posterior (AP) diameter of the dural sac with less than 10mm indicating stenosis (Bolender, Schonstrom, & Spengler, 1985). Verbiest (1975) claimed that spinal stenosis could be assessed by the AP diameter of the canal with greater than 12mm being normal, relative stenosis between 10 and 12mm and absolute stenosis as less than 10mm. It has been suggested to apply Verbiest's initial classification of stenosis to axial CT scans (McCulloch, Transfeldt, 1997). Spengler's group used the space available for the cauda equina as an area in defining stenosis. The area was calculated using the parameters of anterior-posterior (AP) diameter and interpedicle (IP) diameter. They used the values of an area less than 100mm^2 as relative stenosis and an area less than $65\text{-}70\text{mm}^2$ as absolute stenosis. Even more recently a spinal ratio has been introduced as a measurement to help assist in the diagnosis (Laurencin, Lipson, Senatus, Botchwey, Jones, Koris, & Hunter, 1999).

Studies present in the literature do not suggest strong evidence regarding the sensitivity and specificity of these tests in diagnosing spinal stenosis (Kent, Haynor, Larson, & Deyo, 1992). One study assessed CT scan measurements of the osseous canal diameter and postmyelographic CT scan measurements of the dural sac in patients with and without clinical evidence of stenosis. They found that the postmyelographic CT scan was much more accurate than the CT scan alone in predicting which patients had stenosis (Bolender, 1985). However these measurements were made by one observer who was not blinded and as a result this enters a significant bias in measurement. A myelogram is an invasive study and carries risks such as infection, epidural hematoma and spinal headache. If one can avoid such risks by a non-invasive procedure such as a plain CT scan or MRI it would obviously be preferable.

Although the accuracy of CT scans and MRI have been evaluated by comparing preoperative investigations to intraoperative findings, no study has been performed assessing the reliability of such measurements by assessing the degree of interobserver variability (Modic, 1986). Hence the purpose of this study was to assess the degree of interobserver variability when measuring the osseous canal diameter in assessing degenerative lumbar spinal stenosis. A second purpose of the study was to evaluate the overall assessment of the degenerative spine and whether there was agreement between observers in classifying the stenosis as simply mild, moderate or severe.

4.4 Methods

One hundred CT scans of lumbar spinal stenosis as reported by radiology reports were reviewed from the databases between 1995 and 2002. These scans were reviewed by the researcher and twenty five scans were selected. Some of the CT scans were felt to be normal spines despite the suggestion of stenosis by the reporting radiologist these were included in the study. Prior to the start of the study the number of radiographs required was calculated to be 18 in accordance with rules for kappa analysis (Cicchetti, 1977) (Kramer, 1981). Cicchetti (1977), proposed that one could estimate the number of cases needed to determine a kappa by squaring the number of categories and multiplying by four. The CT scans were selected by the author who was not a reviewer and thus independent of the results. Although the CT scans were selected from the institution where the reviewing surgeons practice, the reviewers were blinded as to patients' names, treating surgeons, and the treatments the patients received.

From the selected CT scans the soft tissue windows were selected. From these images, the images which represented L3-4, L4-5, and L5-S1 at the mid pedicle level were marked representing a total of 75 images. The CT scans were then distributed to three orthopaedic surgeons, one senior orthopaedic resident, two radiologists and a neurosurgeon. The participants were also provided with the exact same calibrated ruler for measuring the required distances. After all measurements had been taken the researcher transformed the measurements into real values on the basis of the individual scales of each image. They were asked to measure the spinal canal and record the anterior-posterior (AP) and interpedicular (IP) distances in millimeters of the marked images representing 150 measurements per participant. As well the participants were

asked to classify the spinal canal as mild, moderate, or severe stenotic, or normal if they felt the CT scan did not represent stenosis.

The results were subsequently tabulated. The Verbiest classification was treated as ordinal values as was the classification of mild, moderate or severe stenosis. A weighted kappa was chosen as the analysis of choice for both sets of data. Agreement of the measurements, a continuous variable was performed using an analysis of variance (ANOVA).

4.5 Results

First an analysis of the overall agreement between the six observers using the overall scale of normal, mild, moderate, or severe was performed using a weighted kappa analysis. The weighted kappa achieved was 0.51 with a 95% confidence interval of (0.43, 0.60). A second analysis was performed using this classification to determine the degree of agreement between the three orthopaedic surgeons and a weighted kappa of 0.58 was achieved with a 95% confidence interval of (0.39,0.77). A third analysis was performed to determine the degree of agreement between the two radiologists and a weighted kappa of 0.58 was achieved with a 95% confidence interval of (0.24,0.92).

The data were then re-tabulated and the degree of agreement was assessed for each level of the scan L3-4, L4-5, L5-S1 using Verbiest classification. First, for the image at L3-4 a weighted kappa of 0.25 was achieved with a 95% confidence interval of (0.18, 0.32). A second analysis was performed for the image at L4-5 and resulted in a weighted kappa of 0.22 with a 95% confidence interval of (0.15, 0.29). A final analysis

of Verbiest classification was performed for L5-S1 and a weighted kappa of 0.26 was achieved with a 95% confidence interval of (0.18, 0.34).

An analysis of agreement was performed using ANOVA for the measurements of the AP diameter at the L4-5 level regardless of classification and a p value of < 0.001 was achieved.

Finally, the data were once again re-tabulated on the basis of the mean measurements of the AP diameter at L4-5 for all observers on the basis of how they classified the canal (i.e. normal, mild, moderate, or severe). The results are summarized in Table 10. Four separate ANOVAs were performed on the basis of how they were classified; normal, mild, moderate, or severe and the results are summarized in Tables 11 to 14. First, considering the data on the canals which were classified as normal ANOVA revealed F value of 0.347 and a p-value of 0.843. A second ANOVA was performed for those canals which were classified as mild and an F-value of 2.72 and a p-value of 0.034. A third ANOVA was performed on the canals classified as moderate and revealed an F-value of 8.32 with an associated p-value of 0.000. Finally an ANOVA was performed for the group which were considered severe and an F-value of 5.80 with a p-value of 0.001 was achieved.

Table 10: Mean values for classification of stenosis at L4-5 AP distances for each observer in millimeters.

Observer	Normal	Mild	Moderate	Severe
1	17	14	10	9
2		11	7.6	6.9
3	16.5	14.2	12.9	12.0
4	17	14.3	12.0	11.6
5	16.2	13.4	12.6	13.2
6	18.9	16.9	15.8	14

Table 11: The mean measurements of AP diameter of L4-5 for spines classified as normal, as well as the ANOVA.

Observer	Mean	St Deviation	F	p-value
1	17.00	6.72	0.347	0.843
2	16.50	4.95		
3	17.00	3.06		
4	16.17	1.49		
5	18.90	1.39		
6	17.48	0.91		

Table 12: The mean measurements of AP diameter of L4-5 for spines classified as mild, as well as the ANOVA.

Observer	Mean	St Deviation	F	p-value
1	14.00	3.64	2.72	0.034
2	11.40	3.78		
3	14.17	3.1		
4	14.33	1.37		
5	13.44	1.59		
6	16.89	2.71		

Table 13: The mean measurements of AP diameter of L4-5 for spines classified as moderate, as well as the ANOVA.

Observer	Mean	St Deviation	F	p-value
1	10.00	2.89	8.322	0.000
2	7.58	2.31		
3	12.89	2.02		
4	12.00	2.61		
5	12.60	2.07		
6	15.80	3.83		

Table 14: The mean measurements of AP diameter of L4-5 for spines classified as severe, as well as the ANOVA.

Observer	Mean	St Deviation	F	p-value
1	8.63	2.92	5.799	0.001
2	6.88	1.46		
3	12.00	3.63		
4	11.60	1.82		
5	13.20	2.17		
6	14.00	-		

Chapter 5- Discussion

5.1 Results

5.1.1 Calcaneal Fractures

Like any classification system, in order to facilitate discussion of potential treatment options for a given fracture of the os calcis, it is important that interpretation of the classification be uniform and universal. Sanders' classification attempts to achieve this goal and potentially allows surgeons to alter potential surgical interventions based on the classification. It was our hypothesis that the Sanders classification did in fact achieve this goal and that there was minimal inter-observer variability in using the classification system. The results obtained supported our hypothesis.

It was decided to assess inter-observer variability in terms of both the classes themselves in addition to the subclasses. It was felt that this was essential to evaluate variability in the overall classes, as well to evaluate the variability in using subtypes in order to fully assess Sanders' classification.

The results obtained for the amount of variability between users for the classes as a whole revealed a weighted kappa of 0.48, which represents a moderate strength of agreement (Kramer, 1981). Consequently, one can deduce that there was reasonable agreement among surgeons in classifying os calcis fractures irrespective of subtypes.

The results obtained for the amount of variability between users when considering the classes and the subclasses revealed a weighted kappa of 0.56, which represents again a moderate strength of agreement (Kramer, 1981). Once again it was concluded that there was a reasonable degree of consistency between observers when classifying os calcis fractures using Sanders' classification in terms of both classes and subclasses.

The conclusion that the classification system represented consistency between users was heightened by the fact that the users in this study were not all foot and ankle surgeons, or trauma surgeons. The population of surgeons surveyed included a hand and upper limb surgeon, a foot and ankle surgeon, an arthroplasty/sports medicine surgeon and a senior resident.

The analytic test of choice, the weighted kappa test, is reserved for ordinal values. This therefore assumes that a grade III fracture is worse than a grade II. It was felt that this was a legitimate assumption, however it should be brought to the attention of the reader. Although the same assumption does not apply for the subclasses, that is a grade IIAC is worse than a grade IIAB, a weighted kappa still provides an assessment of the degree of agreement and may in fact provide an underestimate of the degree of agreement. As well the number of observers chosen was relatively small and introducing more observers may tighten the confidence intervals.

Overall, it was concluded that Sanders' classification system proved to achieve moderate agreement among users, thus representing a useful classification system. The level of agreement supports the conclusion that there is consistency and uniformity in the utilization of the classification.

5.1.2 Subtrochanteric Femur Fractures

Subtrochanteric femur fractures are often technically challenging fractures to treat. Currently there is no single agreed upon method to treat these fractures and surgeons are often left with therapeutic dilemmas. A coherent classification system should assist surgeons in their treatment approach to and evaluation of the results of these fractures. A classification should not only group and organize, but should allow communication with a common understanding. If this is achieved the surgeon will be able to utilize the classification to direct assessments and potential treatments. Thus, in any classification of subtrochanteric fractures it is imperative that the classification be validated with regards to variability in its use. Russell and Taylor designed a classification which attempted to achieve this goal and allow surgeons to alter treatment plans according to the classification. It was our hypothesis that the Russell-Taylor classification did in fact achieve this goal. That is, it was suspected that there was minimal interobserver and intraobserver variability in using the classification system. The results obtained did not support this hypothesis.

Given that the essence of the class differentiation is based on the involvement of the piriformis fossa it was decided to assess observer variability in terms of both the classes themselves in addition to the subclasses. When considering the use of an intramedullary device versus a plate and screws to treat this fracture, an important consideration is if the fracture line extends to the piriformis fossa, the entry point for an intramedullary nail. Therefore it was felt that it was essential to consider both the classes as a whole and the subclasses when evaluating observer variability.

The results obtained for the interobserver variability between users for the classes as a whole revealed a weighted kappa of 0.056 at the first set of observations and 0.12 for the second set of observations, which represents a slight to poor strength of agreement and is slightly better than chance alone (Kramer, 1981). An extensive degree of interobserver variability exists when considering the classes as a whole. That is, irrespective of subtypes, there was minimal agreement in determining if the piriformis fossa was involved in the fracture.

When considering the classes and the subclasses together for both sets of observations a weighted kappa of 0.31 was achieved on the first set of data and a weighted kappa of 0.32 was achieved on the second set of observations. Thus representing only a fair degree of agreement between observers (Kramer, 1981). This allows the conclusion that there was a moderate degree of inconsistency between observers when considering extension of the fracture of both the piriformis fossa and the lesser trochanter.

When considering intraobserver variability in the subclasses, a kappa of 0.64 was achieved for observer one. A kappa of 0.28 was achieved for observer two. A kappa of 0.77 was achieved for observer three. A kappa of 0.42 was achieved for observer four. Overall, the observations represent a moderate degree of intraobserver agreement in determining extension of the fracture line to the piriformis fossa and lesser trochanter (Kramer, 1981).

When considering classes as a whole, intraobserver variability revealed a kappa of 0.73 for observer one. A kappa of 0.35 was achieved for observer two. A kappa of 0.88 was achieved for observer three. A kappa 0.77 of was achieved for observer four. Overall, the observations represent a fair degree of intraobserver agreement in determining the extent of the fracture line to the piriformis fossa and to the lesser trochanter. As a consequence overall, intraobserver variability does not appear to be a significant issue in assessing the validity of this classification system, with the exception of observer two, who may be an outlier. The lack of intraobserver consistency for this observer may be due to numerous reasons, some of which are dealt with later in this thesis.

The weighted kappa test is reserved for ordinal values (Kramer, 1981). This assumes that a grade II fracture is worse than a grade I; likewise, a grade IIB is worse than a grade IIA. It was felt that these were legitimate assumptions.

The number of observers chosen, four, was relatively small. Using larger numbers of observers may narrow the confidence interval. The number of radiographs observed was only sixteen, which appears small, however the number of observations needed to be used was calculated prior to starting the investigation. This was based on a formula proposed by Cicchetti (1977) in which sixteen is sufficient to detect the degree of agreement. Cicchetti (1977), proposed that one could estimate the number of cases needed to determine a kappa by squaring the number of categories, in this case two, and multiplying by four.

The observers in this study were unable to strongly agree on whether fractures extended to the piriformis fossa, and consequently the entry point for an intramedullary

nail. This would lead one to question the usefulness of plain radiographs in determining the best course of treatment when considering an intramedullary nail. Perhaps surgeons should consider the use of a preoperative CT scan when considering using an intramedullary nail to ensure the fracture does not extend to the entry point, namely the piriformis fossa fractures. Nailing such a fracture could lead to devastating consequences.

Overall, it was concluded that the Russell-Taylor classification system proved to achieve minimal agreement among and between users, thus representing a less than reliable classification system. The level of agreement supports the conclusion that there is moderate degree of inconsistency in the utilization of the classification and therefore contributes to the assessment of the validity of this classification system.

5.1.3 Spinal Stenosis

Despite the fact that spinal stenosis is often a clinical diagnosis the literature as well as patient charts are often filled with descriptive terms such as mild or moderate stenosis as well as measurements of the canal. No studies have been performed to assess the reliability of any descriptive classification in assessing spinal stenosis. Although spinal stenosis is often diagnosed clinically it is essential to have proper radiographic evidence to support the diagnosis and direct treatment. This radiographic evidence must be accurate and reliable, otherwise it may not only be unhelpful but in fact misleading to the attending physician.

For the results of any diagnostic test to be useful it must be both reliable and valid. This will enable physicians interpreting the tests to communicate with a common understanding. Like any classification system, in order to facilitate discussion of potential treatment options for spinal stenosis, it is important that interpretation of investigative tests as well as the subsequent classification be reliable and valid.

The purpose of this study was to assess the degree of reliability of assessing spinal stenosis using CT scans. It was our hypothesis that using either the Verbiest classification of stenosis based on measurements, or by simply classifying the canal as normal, mild, moderate, or severe that there would be a high degree of interobserver reliability. The results modestly supported this hypothesis. As well it was hypothesized that there would be a strong degree of agreement between observers when measuring the diameters of the canal directly. The results failed to support this. Finally, it was hypothesized that there would be interobserver reliability between the measurements of the canal space and the classification of the canal as normal, mild, moderate, or severely stenotic. The results failed to support this hypothesis.

The results obtained for the amount of variability between all observers classifying the scan as a whole as either normal, mild, moderate, or severe stenosis revealed a weighted kappa of 0.51, which represents a moderate strength of agreement (Kramer, 1981). Consequently one can deduce that there was reasonable agreement among observers in classifying the degree of stenosis for the spinal canal overall, irrespective of the level involved.

The results obtained for the amount of variability between orthopaedic surgeons and radiologists was assessed independent of each other classifying the scan as a whole

as either normal, mild, moderate, or severe stenosis. Orthopaedic surgeons as a group achieved a weighted kappa of 0.58, which represents a moderate strength of agreement (Kramer, 1981). Radiologists as a group also achieved a weighted kappa of 0.58. Consequently one can deduce that there was reasonable agreement among both orthopedic surgeons and radiologists in classifying the degree of stenosis for the spinal canal overall, irrespective of the level.

Next the results obtained for the amount of variability between users when considering Verbiest's classification (1975) based on AP measurements of the canal was considered. He classified the canal as either normal ($>12\text{mm}$), relative stenosis ($10\text{--}12\text{mm}$), or absolute stenosis ($<10\text{mm}$). Using these criteria for the three levels of the lumbar spine investigated, weighted kappa analyses were performed. For the L3-4 level a weighted kappa of 0.25 was achieved, for L4-5 a weighted kappa of 0.22 was achieved, and for L5-S1 a weighted kappa of 0.36 was achieved. These results represent only a fair degree of agreement (Kramer).

In comparing the results using two separate but similar classification systems there are significantly different weighted kappas achieved. That is, 0.51 for the spine as a whole versus 0.22-0.36 for individual levels. Reasons for this may be that different observers felt different levels were classified differently but overall they felt the canal was the same. Or it may be that different observers had criteria, subjective or objective which were not in keeping with the measurements originally described by Veibiest (1975).

Therefore, an analysis of the measurements versus the classification was made on the most common level of spinal stenosis L4-5 (McCulloch, 1993). Although there was

moderate agreement classifying the spine as a whole, and fair agreement in classifying this individual level, the measurements between different observers varied immensely, both from each other and within themselves as to how they classified the canal overall. For canals which were classified as normal the ANOVA revealed a p-value of 0.843 therefore it can be suggested that the mean measurements of the AP canal diameter in the group of scans the observers classified as normal were not statistically different, that is there was agreement in the measurements of normal canals. There is of course the possibility that this result represents a type II error. That is, it is possible that we have concluded that no difference between the groups existed when in fact it did, the probability that this has occurred is represented by the term Beta. The power of a study refers to the probability of concluding that there was a difference between the two groups when in fact there was one. This concept is closely related to type II error. As the power of a study increases the chances of committing a type II error diminishes. Therefore in order to confirm the above conclusions a power analysis would have to be performed.

In canals which were classified by the observers as mild the ANOVA revealed a p-value of 0.034. One can conclude that there were statistically significant differences in the measurements of the AP diameter of the canal in those spines that were classified as mild stenosis by the observer. Likewise ANOVAs for the scans which were considered moderate and severe stenosis revealed p-values of <0.001 and 0.001 respectively. Therefore one can conclude that there were statistically significant differences in the measurements of AP diameter of the canal for those which were classified as moderate or severely stenotic.

As a result of the ANOVA analysis at L4-5 some important observations and conclusions can be drawn. Even though there was a moderate degree of agreement between observers in assessing if the canal was mildly, moderately or severely stenotic, the measurements used to determine this reflect no degree of agreement. As well, there is agreement both in the classification of normal spines as evidenced by the kappa value, as well as agreement in measurements of the canal diameter in these individuals.

The conclusions of these results is heightened by the fact that the users were not all spine, or neurosurgical subspecialists. The three orthopaedic surgeons were as follows: a foot and ankle surgeon, a senior orthopaedic resident, and a spine surgeon. The neurosurgeon had fellowship training in spine surgery. The two radiologists were as follows: a general radiologist and a musculoskeletal radiologist.

Overall, it was concluded that classifying stenosis of the lumbar spine using a simple descriptive classification system of mild, moderate, or severe proved to achieve moderate agreement among users, and between specialties thus representing a useful classification system. The level of agreement supports the conclusion that there is reliability and uniformity in the utilization of this classification. The level of agreement was somewhat less when considering the exact measurements either through the classification of Verbiest or through the overall measurements of the canal.

5.2 Statistics

Diagnostic tests such as radiographs, CT scans, EKG, blood work and mammography are used to direct a physician's diagnosis and treatment of disease. Whenever these tests involve interpretation by humans they are subject to error. When assessing a diagnostic test we often compare it to what it is supposed to assess, that is, does it reflect the gold standard (Sackett, 1991). For example, does a positive mammogram actually reflect the gold standard, pathology. The real question is whether or not the test is valid. This introduces terms such as reliability, accuracy and bias.

The examples in this thesis reflect the degree of agreement between users in interpreting the results of the same test, or the degree of agreement between the same user at different times in interpreting the results of the same test. This reflects the terms interobserver and intraobserver reliability or consistency. Therefore, a general definition of reliability could be the extent to which examinations of the same patient or specimen agree with one another (Sackett, 1991).

Validity is also influenced by the degree of accuracy in the test, that is the closeness of a clinical observation to the true clinical state. For example, in showing medical students wrist x-rays and asking them if there was a scapholunate injury, they may all say no, and therefore have a high degree of interobserver reliability. However, they also may be wrong and therefore inaccurate. Inaccuracy may also be influenced by bias, that is, a systematic deviation of an observation from the true clinical state.

Validity may be assessed as internal or external validity. External validity refers to how well the test actually reflects reality. Internal validity refers to how well the actual measurements represent the variables of interest (Hulley, & Cummings, 1988). In order for a test to be valid it implies that the test is close enough to the truth to make it useful (Jeaeschke, Guyatt, & Sackett, 1994). The validity of a test is influenced by both sampling error and measurement error, each of which has random and systematic errors (Hulley, & Cummings, 1988). To enhance the test's validity, the test must be precise and free of random error, as well as accurate and free of systematic errors. In order to improve the validity of a test both precision or reliability and accuracy of these factors need to be assessed and enhanced. A measurement or test may be reliable and accurate but not valid, but it cannot be valid without being reliable and accurate. That is, reliability and accuracy are necessary but not sufficient conditions for validity.

Further concepts surrounding validity which deserve mention at this point include criteria and construct validity (Sackett, 1991). Criterion related validity demonstrates the accuracy of a measure or procedure by comparing it with another measure or procedure which has been demonstrated to be valid, that is a gold standard. Orthopaedic surgery often lacks a true gold standard, and the surgeon has to rely on a further validity concept, construct validity.

Construct validity seeks agreement between a theoretical concept and a specific measuring device or procedure (Sackett, 1991). For example, a classification of a fracture is often a theoretical concept based on radiographs which then direct the specific

treatment modalities. Orthopaedic surgery often lacking a validated gold standard relies more heavily on the construct validity.

To evaluate construct validity, steps should be followed. Theoretical relationships must be specified, the empirical relationships between the measures of the concepts must be examined, finally the empirical evidence must be interpreted in terms of how it clarifies the construct validity of the particular measure being tested (Carmines & Zeller, 1991).

When assessing the validity of a proposed test, such as a classification system, there are several questions which must be asked. First, has there been a blind comparison to a gold standard (Sackett, 1991) (Jeaeschke, 1994)? How well does the test reflect what it is supposed to, in other words how accurate is the test? In the case of traumatic and nontraumatic orthopaedic conditions the gold standard is usually intraoperative findings. When used simply to confirm the pathology, diagnostic imaging can be easily compared to the gold standard. For example, an x-ray determines there is a fracture and this is confirmed by surgery. However when the x-ray is used to determine the severity of the pathology in comparison to surgery it becomes more complex (Sackett, 1991), and this analysis is usually not performed on orthopaedic classification systems.

Next it must be determined if the test has been evaluated in a sample of patients, with the appropriate spectrum of disease, with individuals with different but commonly confused pathologies (Sackett, 1991). The key value of a diagnostic test or classification system is its ability to distinguish pathologies which are often confused, especially if their

prognosis and treatment options are different. For example, a fracture which is heavily comminuted is not often confused with one which is undisplaced. However, whether the fracture has two parts, three parts or four parts can be confusing.

Next, one must consider the setting in which the test was originally described. Does the setting represent a group of patients reflective of the population? For example, if a classification system was developed in a level one trauma center who only saw polytraumatized patients, its applicability to the community hospital may be limited. Most of the literature concerning classifications do not address this issue.

Another consideration is whether or not the reliability of the test been examined. That is how precise is the measurement or test (Sackett, 1991).? In order for a test or classification to be useful it must be reproducible and observer variability must be minimized. A significant portion of classifications and tests in orthopaedics have not examined this issue.

Also, it must be determined if the description of the test has been adequate to provide for replication. Have the proponents of the test or classification provided adequate information to allow the test to be replicated (Sackett, 1991)? For example, when considering classifications of fractures, have they adequately described what views they based their classification on? Most orthopaedic classifications do describe what views they are based on and how to achieve those views.

Finally, has the utility of the test been determined? That is, is a patient better off having had this test (Sackett, 1991)? This may be applied to the use of CT classifications

in orthopaedics. For example, should a patient with a plain x-ray finding of a calcaneal fracture have a CT scan to help classify and prognosticate the injury? In deciding this, one must consider patient factors, individual fracture factors and cost. Most classifications of traumatic fractures will attempt to do this based on the prognosis provided through classification.

Overall, using the above questions as a guide, the validity of a test demands the absence of both systematic bias and random bias. That is, it should be precise and accurate (Sackett, 1991). These two terms warrant further discussion.

Precision refers to how well a diagnostic test will give the same result when assessing the same measurement repeatedly. That is, a precise test will make the same conclusions each time it is used. For example, a well calibrated scale will be very precise in measuring weights, however a subjective interpretation of radiographs may not be so precise. Precision is closely linked to and interchangeable with the terms reliability and consistency. It is influenced mainly by random error. Therefore, in order to improve a diagnostic test one must search for and eliminate the sources of error. The three main sources of error are as follows: observer variability (e.g. how an observer uses the test), subject variability (e.g. inherent differences in patients anatomy), and instrument variability (e.g. differences in X-ray technique) (Hulley, & Cummings, 1988).

Precision can be assessed based on statistical analysis through standard deviations, kappa analysis or coefficient of variation (Kramer, 1981). Several descriptions of this have been made. Test-retest reliability refers to concordance of repeated measures of the

same sample at different times (Hulley, 1988). Internal reliability assesses concordance between how different tests assess the same outcome. Interobserver and intraobserver reliability refers to consistency of measurements both between observers and for repeated measurements of an individual observer at different times.

Hulley and Cummings (1988) listed five ways to eliminate these errors and thereby improved the reliability of a diagnostic test. First, using strict operational definitions on how to use a test or measurement would better standardize the results. For example, if stricter criteria were used for a classification of fractures. Training observers in how to use a measurement or test could theoretically reduce error. Refining instruments, whether it be a ruler or clarifying questionnaires, should reduce error. When direct measurements using an instrument are involved, automating the instrument reduces human error. Finally, precision of a test could be enhanced by repetition thereby decreasing the influence of random error. Applying these rules to a diagnostic test could enhance precision thereby leading to a more valid test. A test may be entirely precise, however if it is not measuring what it is supposed to it will not be accurate and therefore not valid.

Accuracy refers to the extent to which a test result represents the truth. Accuracy is not necessarily related to precision. For example, a test of a patient's blood count after trauma may be very precise as the same levels were achieved on multiple tests. However, if the patient has received excessive fluids, the samples may be diluted and therefore not reflect the true blood count, and are therefore inaccurate. Although not linked, the

concepts of accuracy and precision will often go together and strategies at improving accuracy will often improve precision and vice versa (Hulley, 1988).

Accuracy is subject to systematic bias. Similar to precision there are three main influences of bias. First, observer bias refers to a consistent distortion of the reporting of results by the observer. For example, a surgeon classifying a fracture knowing the results of subsequent treatment (Hulley, 1988). Second, subject bias refers to distortion of measurements by the study subject (Hulley, 1988). For example, a patient with back pain may report the severity of their pain influenced by such issues as litigation or workers compensation. Third, instrument bias, refers to consistent error within the machine. For example, if the CT gantry was consistently mal-positioned during a scan it may alter the interpretation of results.

The accuracy of a test can be assessed by directly comparing it to a gold standard. However, as alluded to earlier this is not as straightforward as it sounds in orthopaedic surgery. Means of improving accuracy include the ones previously discussed for precision as well as the following. The first is making an unobtrusive measurement (Hulley, 1988). For example, in assessing patients with back pain and injury, following the patients without them knowing would provide a means of assessing their level of activity without them being aware, thereby reducing bias. Next, blinding the observers of a test or classification would reduce the bias (Hulley, 1988). For example blinding the assessors of a classification of a fracture as to what treatments and outcomes patients had would reduce the bias. Finally, calibrating instruments against a gold standard would reduce error (Hulley, 1988).

Overall, to enhance the validity of a diagnostic test, one must consider the precision and accuracy of the test. In order for a test to be valid it must be both reliable and accurate. Strategies at improving these have been outlined. The research in this thesis largely focused on the reliability of classifications. Given that not all the guidelines for assessing validity of a diagnostic test can be met for fracture classifications, such as accuracy (e.g. comparing with a gold standard), it is essential that the remaining guidelines, such as assessment of reliability and precision be scrutinized. It is with this goal in mind that the three projects of this thesis were performed.

Assessing the reliability in using Sanders' classification of calcaneal fractures, Russell-Taylor classification of subtrochanteric fractures, and measurements of canal diameter in patients with lumbar spinal stenosis, adds to the assessment of the validity of these classification systems. Reviewing the literature there were no data available with regards to the accuracy, or overall validity of any of these classifications systems. Therefore by assessing reliability one is able to assess at least one of the components of validity. By satisfying one component of validity, reliability, surgeons are able to interpret the literature with more confidence regarding treatment and prognosis. For example Buckley et. al. (2002) suggested that surgical treatment of Sanders type II calcaneal fractures had better results than types III and IV treated surgically. Prior to the study in this paper interpreting Buckley's results could be questioned on the basis of classification. That is, were the results real and valid or were they biased by the fact that there was no reliability between those who used Sanders classification in classifying the fractures preoperatively.

5.2.1 Analysis of Agreement

Research may be concerned both with trends and concordance, or agreement between variables. Trends refer to the strength of changes in one variable to affect the other. Evaluation of trends utilizes indices such as the Pearson correlation coefficient, regression coefficient, Spearman's ratio, or Kendall's tau (Kramer, 1981). Other than trends studies may be concerned with concordance, that is the extent to which one variable can replace the other, or the extent to which the variables yield the same results (Kramer, 1981). The measurements of trends cannot be applied to concordance.

A factor that one must consider when assessing concordance is the type of variable being assessed. Data may be nominal, that is categorical without order, or, ordinal, that is categorical with an implied rank order. A dichotomous variable refers to a variable which only has two options, therefore agreement is either all or none. When assessing the degree of agreement or disagreement between nominal variables one can adapt an approach similar to the one for dichotomous variables (Kramer, 1981). For ordinal variables the variable has three or more ordered options and agreement may not be all or none: for example, mild, moderate, or severe. Therefore, when assessing agreement, there is a level of disagreement. For example, if one observer said the variable in question was mild while the other said it was severe, this represents more disagreement than if they said it was mild and moderate (Kramer, 1981). One way of dealing with this discrepancy is to assign weights to the amount of disagreement. For total agreement would receive a weight of 0, if one observer said mild and the other moderate it would receive a weight of 1, and if one observer said mild and the other

severe it would receive a weight of 2. Conversely, one could do a similar calculation weighting the agreement where a score of 0 would represent maximum disagreement (Kramer, 1981).

The role of chance is not insignificant when assessing the degree of agreement between observers. Whenever two outcomes are compared there is a finite probability that they will agree entirely by chance. Just as one has a 25% chance of guessing a multiple choice question when there are four options, chance plays a role in assessing concordance. Consideration of chance is especially important when the number of categories is small.

There are several proposed ways of assessing observer variability. Originally, researchers used percent overall agreement to assess agreement. How this statistic is arrived at is best illustrated through an example. Two surgeons were asked to classify fractures as normal, undisplaced, displaced or comminuted. The results are cross tabulated and an overall percent agreement can be calculated by adding the number of times the observers agreed and dividing the sum by the total number of cases (Gordis, 2000). This measurement can be easily skewed however. For example, there is usually little disagreement between observers when the x-ray is normal. Therefore, if the sample of cases has a large portion of normal x-rays the percent agreement would be high and may not truly reflect the degree of agreement amongst observers and in fact may skew the results to show significant agreement when the observers may have absolutely no agreement on difficult fractures. (Gordis, 2000)

In attempting to eliminate this bias statisticians have suggested that removal of the subjects who were deemed normal would reduce the bias. A calculation is then

performed of the percent agreement using a denominator which represents only observations which were labeled abnormal by at least one observer (Gordis, 2000).

Another approach to assess agreement is to treat categorical data as if they were actually interval data. For example, if a fracture is undisplaced it is given a value of 0 and if it is displaced it is given a value of 1. So, if there were 200 fractures, there would be 200 sets of values with one of the following combinations (0,0), (1,0), (0,1), or (1,1). Using these paired data one could use a correlation coefficient (Norman, & Streiner, 1994). Using the Pearson correlation coefficient and using mathematical manipulations the phi coefficient of correlation can be derived.

The researcher must then decide, as with any correlation coefficient, if the degree of correlation is adequate.

There are limitations to both of these statistics. Namely, the correlation coefficient approach ignores systematic observer bias. For example, if observer 1 consistently ranks observations higher than observer two, who is also consistently using the same order rankings, then there may be a high degree of correlation but absolutely no agreement. The percent agreement does not account for agreement by chance alone (Gordis, 2000). That is, there will be a certain amount of agreement amongst observers regardless of the criteria they use to classify the fractures. For example, if you asked a first year class of medical students to classify a set of fractures as I, II or III there would be a certain amount of agreement by chance. Cohen (1960) recognized this limitation and developed a new statistic based on eliminating the amount of agreement based on chance, thereby reflecting the true agreement, the kappa statistic.

5.3 Conclusion

Orthopaedic surgery often relies on classification systems to help direct the treatment of patients and their conditions, and assist in reporting results of these treatments. There are of course other factors which may effect the outcome of a patient, these include such factors as patient compliance, associated comorbidity, and individual surgeon technical abilities. However, the purpose of this thesis was to focus on the step of classification in patient management, how it works, and how it can be improved in order to better patient care.

Given that it is difficult to satisfy all the criteria of validating a diagnostic test for classification systems in orthopaedic surgery it is paramount that the criteria which are able to be assessed be carefully scrutinized. Criteria such as reliability must be assessed in order to at least partially assess the validity of the classification system. In order for the classification system to be valid it must be reliable and accurate. If the classification is not reliable, it cannot be valid. This fact represents the underlying premise of this paper. The purpose of this paper was not to assess validity of the respective classifications, but to assess one component of validity, reliability.

The findings presented are similar to previous studies in the orthopaedic literature. The assessment of the degree of reliability in assessing three separate classification systems lead to only a moderate to fair degree of reliability, thereby questioning the validity of these classification systems.

The lack of validity of classification leads to confusion within the orthopaedic literature and its interpretation. The lack of validity may in fact lead a treating surgeon to prescribe the wrong treatment based on literature which recommends treatments based on classification. This may not be due to poor technique or the surgeon's mistreatment of the condition, but in fact a misclassification of the condition.

Despite problems associated with classification systems they remain an integral part of orthopaedics, if used appropriately. They help guide treatment and serve as an important research tool. As other authors have concluded, we also recommend that in order for a classification to be used in the literature it should be validated, by at least assessing reliability. If a new classification is to be developed it should be carefully scrutinized prior to publication in order to prevent the problems of using an invalid classification system.

References

- Amundsen T, Weber H, Lilleas F, Nordal HJ, Abdelnoor M, Magneas B. Lumbar spinal stenosis: Clinical and radiologic features. *Spine* 1995; 20:1178-1186.
- Andersen DJ, Blair WF, Steyers CM, et al Classification of distal radius fractures: an analysis of interobserver reliability and intraobserver reproducibility. *J Hand Surg* 1996; 21(A): 574-582.
- Andersen E, Jorgensen LG, Hededam LT, Evans' classification of trochanteric fractures: an assessment of interobserver and intraobserver reliability. *Injury* 1990; 21:377-378.
- Andersen GR, Rasmussen JB, Dahl B. et al. Older's classification of Colles' fractures: good intraobserver and interobserver reproducibility in 185 cases. *Acta Orthop Scand* 1991; 62:463-464.
- Anonymous. Fracture and dislocation compendium. Orthopedic Trauma Association Committee for Coding and Classification. *J Orthop Trauma* 1996; 10 (supp 1): v-ix, 1-154.
- Atlas SJ, Deyo RA, Keller RB et al. Maine Lumbar Spine Study, Part III: 1 year outcomes of surgical and nonsurgical management of lumbar spinal stenosis. *Spine* 1996; 21:1787-1795.
- Bernstein J, Adler LM, Blank JE et al. Evaluation of the Neer shoulder system of classification of proximal humerus fractures with computerized tomographic scans and plain radiographs. *J Bone Joint Surg* 1996; 78A: 1371-1375.
- Boden SD, Davis DO, Dina TS, Patronas NJ, Wiesel SW. Abnormal magnetic-resonance scans of the lumbar spine in asymptomatic subjects: A prospective investigation. *J Bone Joint Surg.* 1990; 72 (A):403-408.
- Bolender NR, Schonstrom NSR, Spengler DM. Role of computed tomography and myelography in the diagnosis of spinal stenosis. *J Bone Joint Surg.* 1985; 67-A:240-246.
- Boyd H, Griffin LL. Classification and treatment of trochanteric fractures. *J Trauma.* 1992; 32:71-76.
- Brien H, Nofall F, MacMaster S et al. Neer's classification system: a critical appraisal. *J Trauma* 1995; 38: 257-260.

- Brady OH, Garbuz DS, Mari BA, et al. The reliability and validity of the Vancouver classification of femoral fractures after hip replacement. *J Arthroplasty*. 2000; 15:59-62.
- Brumback RJ, Jones AL. Interobserver agreement in the classification of open fractures of the tibia. *J Bone Joint Surg* 1994; 76A:1162-1166.
- Buckley R, Tough S, McCormac R, Pate G, Leighton R, Petrie D, Galpin R. Operative compared with nonoperative treatment of displaced intra-articular calcaneal fractures. *J Bone Joint Surg*. 2002; 84(A):1733-1744.
- Campbell DG, Garbuz DS, Masri BA et al. Reliability of acetabular bone defect classification systems in revision total hip arthroplasty. *J Arthroplasty*. 2001; 16:83-86.
- Carmines EG, Zeller RA. Reliability and validity assessment. Newbury Park: Sage Publications, 1991: p23.
- Chan PS, Klimkiewicz JJ, Luchetti WT, et al. Impact of CT scan on treatment plan and fracture classification of tibial plateau fractures. *J Orthop Trauma*. 1997; 11:484-489.
- Codman EA, The shoulder. Malabar, FL: Kreiger, 1934.
- Cohen J. A coefficient of agreement for nominal series. *Educ Psychol Meas*. 1960; 20:37-46.
- Colton CL. Telling the bones [editorial]. *J Bone Joint Surg* 1991; 73B: 362-364.
- Cicchetti DV. Testing the normal approximation and minimal sample size requirements of weighted kappa when number of categories is large. *Applied Psychological Measurement* 5 1977: 101-104.
- Craig WL III, Dirschl DR. Effects of binary decision making on the classification of fractures of the ankle. *J Orthop Trauma* 1998; 12:280-283.
- Cummings RJ, Loveless EA, Campbell J, Samelson S, Maruz JM. Interobserver reliability and intraobserver reproducibility of the system of King et al for the classification of adolescent idiopathic scoliosis. *J Bone Joint Surg*. 1998; 80(A): 1107-1111.
- De Villiers PD, Booysen EL. Fibrous spinal stenosis: A report on 850 myelograms with a water soluble contrast medium. *Clin Orthop* 1976; 115: 140-144.

- Dehne E. Fractures of the upper end of the humerus: a classification based on etiology of trauma. *Surg Clin North Am* . 1945; 25:28-47.
- Dirschl DR, Adams GL. A critical assessment of factors influencing reliability in the classification of fractures, using fractures of the tibial plafond as a model. *J Orthop Trauma*. 1997; 11(7): 471-476.
- Dutton KE, Jones TJ, Slinger BS, Scull ER, O'Connor J. Reliability of the Cobb angle index derived by traditional and computer assisted methods. *Australas Phys Eng Sci Med*. 1989 Mar; 12(1): 16-23.
- Garbuz DS, Masri BA, Esdaile J , et al Classification Systems in Orthopedics. *Journal of the American Academy of Orthopaedic Surgery*. 2002; 10: 290-297.
- Garden RS, Low angle fixation in fractures of the femoral neck. *J Bone Joint Surg*. 1968; 50(B): 562-569.
- Gehrchen PM, Neilsen JO, Olesen B et al. Seinsheimer's classification of subtrochanteric fractures- poor reproducibility of 4 observers' evaluation of 50 cases. *Acta Orthop Scand* 1997; 68 (6):524-526.
- Giachino, AA. Uthoff HK. Intra-articular fractures of the calcaneus. *J Bone Joint Surg* 1989; 71(A): 784-787.
- Gordis, L. Epidemiology. W.B. Saunders Philadelphia 2000. 75-80
- Greenfield LJ, Mulholland MW, Oldham KT, Zelenock GB, Lillemoe KD, Surgery: Scientific Principles and Practice. Third ed. Philadelphia etc: Lippincott-Williams & Wilkins Publishers, 2001: 278.
- Groves EWH. Ununited fractures with special reference to gunshot injuries and the use of bone graft. *J Bone Joint Surg*. 1918; 6(B):203.
- Gustilo RB, Andersen JT. Prevention of infection in treatment of one thousand and twenty-five open fractures of long bones. Retrospective and prospective analyses. *J Bone Joint Surg*. 1976; 58(A): 453-458.
- Haddad FS, Masri Ba, Garbuz DS, Duncan CP. Femoral bone loss in total hip arthroplasty: Classification and preoperative planning. *J Bone Joint Surg* . 1999; 81(A):1483-1498.
- Harty M. Anatomic considerations in injuries of the calcaneus. *Orthop Clin North Am* 1973; 4:179-183.

- Hilibrand AS, Rand N. Degenerative lumbar stenosis: Diagnosis and management. *J Am Acad Orthop Surg* 1999; 7:239-249.
- Hoglund EJ. New intramedullary bone implant surgery. *Gynecol Obstet*. 1917; 34:243.
- Hulley SB, Cummings SR. Designing clinical research: An epidemiologic approach. Williams and Wilkins, Baltimore. 1988: 31-42.
- Jaeschke R, Guyatt G, Sackett DL. Users' guide to medical literature: How to use an article about a diagnostic test. *JAMA* 1994; 271(5):389-391.
- Kent DL, Haynor DR, Larson EB and Deyo RA. Diagnosis of lumbar spinal stenosis in adults: A metaanalysis of accuracy of CT, MR and myelography. 1992; *A.J.R* 158:1135-1144.
- Kramer MS, Feinstein, AR. The Biostatistics of Concordance. *Clin Pharmacol Ther*. 1981; 29: 111-123.
- Kundel K, Funk E, Brutscher M, et al Calcaneal fractures: operative versus nonoperative treatment. *J Trauma* 1964; 4:15-56.
- Kreder HJ, Hanel DP, McKee M, et al. Consistency of AO fracture classification for the distal radius. *J Bone Joint Surg*. 1996; 78(B):726-731.
- Kristiansen B, Andersen UL, Olsen Ca, et al. The Neer classification of fractures of the proximal humerus. An assessment of interobserver variation. *Skeletal Radiol*. 1988; 17(6): 420-422.
- Landis RJ, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159-174.
- Laurencin CT, Lipson SJ, Senatus P, Botchwey E, Jones TR, Koris M, Hunter J. The stenosis ratio: A new tool for the diagnosis of degenerative spinal stenosis. *Int Journal of Surg Inter* 1999. 1;2: 127-131.
- Lenke LG, Betz RR, Bridwell KH, et al. Intraobserver and interobserver reliability of the classification of thoracic adolescent idiopathic scoliosis. *J Bone Joint Surg*. 1998; 80(A):1097-1106.
- Management of back pain. In: Porter, RW ed. Edinburgh, Scotland: Churchill Livingstone; 1993, pp59-72.
- Martin JS, Marsh JL. Current classification of fractures. Rationale and utility. *Radiol Clin North Am*. 1997; 35(3):491-506.

- Martin JS, Marsh JL, Bonar SK et al. Assessment of AO/ASIF fracture classification for the distal tibia. *J Orthop Trauma* 1997;11:477-483.
- McCaskie AW, Brown AR, Thompson JR, Gregg PJ. Radiological evaluation of the interfaces after cemented total hip replacement: Interobserver and intraobserver agreement. *J Bone Joint Surg.* 1996; 78(B):191-194.
- McCulloch J, Transfeldt E. McNab's Backache 3rd Ed. Williams and Wilkins, Baltimore 1997. pp 609-663.
- Modic MT, Masaryk T, Boumpfrey F, Goormastic M, Bell G, Lumbar herniated disc disease and canal stenosis: Prospective evaluation by surface coil MR, CT and myelography. *A.J.R. Am J Roentgenol* 1986.; 147:757-765.
- Muller ME, Nazarian S, Koch P, Schatzker J. The comprehensive classification of fractures of long bones. Berlin: Springer-Verlag, 1990.
- Myerson M, Quill G. Late Complications of Fractures of the Calcaneus. *J Bone Joint Surg [Am]* 1993;75(3): 331-41.
- Neer CS II. Displaced proximal humeral fractures: Classification and evaluation. *J Bone Joint Surg.* 1970; 52(A): 1077-1089.
- Neer CS II. Letter to the editor. *J Bone Joint Surg.* 1994; 76(a):789.
- Norman GR, Streiner DL. Biostatistics: The bare essentials. St. Luis. Mosby. 1994.
- Pott P. Remarks on fractures and dislocations: Chirurgical works of Pott. London: Wood & Innes, 325-329. Reprinted in Rang M. Anthology of Orthopaedics. New York: Churchill Livingstone, 1966.
- Robertson GH, Llewellyn HJ, Taveras JM: The narrow lumbar spinal canal syndrome. *Radiology* 1973; 107: 89-97.
- Rockwood C. A., Green D. P., Bucholz R. W. *Rockwood and Green's Fractures in Adults*. Fourth ed. Philadelphia etc: Lippincott-Raven Publishers, 1998.
- Russell-Taylor Classification of Subtrochanteric Fractures. *Skeletal Trauma* 1998: Vol 2: 1891-1897.
- Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical epidemiology: A basic science for clinical medicine. Boston etc: Lippincott-Raven. 1991:51-153.

Sanders R, Regazzonni P. Treatment of subtrochanteric fractures using dynamic condylar screws. *J Orthop Trauma*. 1989; 3(3):206-213.

Sanders R. Intra-Articular Fractures of the Calcaneus: Present State of the Art. *J Orthop Trauma*. 1992; 6(2): 252-265.

Sanders R, Fortin P, DiPasquale T, Walling A. Operative Treatment in 120 Displaced Intra-articular Calcaneal Fractures; Results Using a Prognostic Computed Tomography Scan Classification. *Clin Orthop Related Res*. 1993; 290:87-95.

Sanders R. The Problem with Apples and Organges. *J Orthop Trauma*. 1997; 11(7): 465-466.

Schonstrom HSR, Bolender NR, Spengler DM. The pathomorphology of spinal stenosis as seen on CT scans of the lumbar spine. *Spine* 1985; 10:806-811.

Seinsheimer F III. Subtrochanteric fractures of the femur. *J Bone Joint Surg*. 1978; 60 A: 300-306.

Sidor ML, Zuckerman JD, Lyon T et al. The Neer classification system for proximal humerus fractures: An assessment of interobserver reliability and intraobserver reproducibility. *J Bone Joint Surg*. 1993; 75(A): 1751-1755.

Siebenrock KA, Gerber C. The reproducibility of classification of fractures of the proximal end of the humerus. *J Bone Joint Surg*. 1993;75(A): 1745-1750.

Soeken KL, Prescott PA. Issues in the use of kappa to estimate reliability. *Medical Care*, 24: 733-741.

Swiontkowski MF, Sands AK, Agel J, Diab M, Kreder H. Interobserver variation in the AO/OTA fracture classification system for pilon fractures: Is there a problem? *J Orthop Trauma*. 1997; 11(7): 467-470.

Taylor JC, Russell TA, LaVelle DG, et al. Clinical results of 100 femoral shaft fractures treated with Russell-Taylor interlocking nail system. *Orthop Trans* 1987; 11:471 (abst).

Teasdale G, Jennett B. Assessment of coma and impaired consciousness: a practical scale. *Lancet*. 1974; 2:81-84.

Thomsen NOB, Overgaard S, Olsen LH, et al. Observer variation in the radiographic classification of ankle fractures. *J Bone Joint Surg*. 1991; 73(B):676-678.

Verbiest H. Pathomorphologic aspects of developmental lumbar stenosis. *Orthop Clin North Am* 1975; 6:177-196.

White AA III, Panjabi MM, Southwick WO. Biomechnaical analysis of clinical stability in the cervical spine. *Clin Orthop* 1975. 109:89.

Wright JG, Feinstein AR. Improving the reliability of orthopaedic measurements. *J Bone Joint Surg.* 1992; 74(B): 287-291.

Appendix A

**Memorial University Of Newfoundland
Faculty of Medicine
Research Project Questionnaire**

The questionnaire below will take approximately 20-30 minutes to complete. Could you please examine each of the CT Scans and classify the calcaneal fracture according to Sanders classification. Thank you in advance for your time.

CT #1

Type I ☐
Type II A ☐
Type II B ☐
Type II C ☐

Type III AB ☐
Type III AC ☐
Type III BC ☐
Type IV ☐

Quality of the CT Scan Poor ☐ Fair ☐ Excellent ☐

CT #2

Type I ☐
Type II A ☐
Type II B ☐
Type II C ☐

Type III AB ☐
Type III AC ☐
Type III BC ☐
Type IV ☐

Quality of the CT Scan Poor ☐ Fair ☐ Excellent ☐

CT #3

Type I ☐
Type II A ☐
Type II B ☐
Type II C ☐

Type III AB ☐
Type III AC ☐
Type III BC ☐
Type IV ☐

Quality of the CT Scan Poor ☐ Fair ☐ Excellent ☐

CT #4

Type I ☐
Type II A ☐
Type II B ☐
Type II C ☐

Type III AB ☐
Type III AC ☐
Type III BC ☐
Type IV ☐

Quality of the CT Scan Poor ☐ Fair ☐ Excellent ☐

CT #5Type I ☐Type II A ☐Type II B ☐Type II C ☐Type III AB ☐Type III AC ☐Type III BC ☐Type IV ☐

Quality of the CT Scan

Poor ☐Fair ☐Excellent ☐**CT #6**Type I ☐Type II A ☐Type II B ☐Type II C ☐Type III AB ☐Type III AC ☐Type III BC ☐Type IV ☐

Quality of the CT Scan

Poor ☐Fair ☐Excellent ☐**CT #7**Type I ☐Type II A ☐Type II B ☐Type II C ☐Type III AB ☐Type III AC ☐Type III BC ☐Type IV ☐

Quality of the CT Scan

Poor ☐Fair ☐Excellent ☐**CT #8**Type I ☐Type II A ☐Type II B ☐Type II C ☐Type III AB ☐Type III AC ☐Type III BC ☐Type IV ☐

Quality of the CT Scan

Poor ☐Fair ☐Excellent ☐

CT #9Type I ☐Type II A ☐Type II B ☐Type II C ☐Type III AB ☐Type III AC ☐Type III BC ☐Type IV ☐

Quality of the CT Scan

Poor ☐Fair ☐Excellent ☐**CT #10**Type I ☐Type II A ☐Type II B ☐Type II C ☐Type III AB ☐Type III AC ☐Type III BC ☐Type IV ☐

Quality of the CT Scan

Poor ☐Fair ☐Excellent ☐**CT #11**Type I ☐Type II A ☐Type II B ☐Type II C ☐Type III AB ☐Type III AC ☐Type III BC ☐Type IV ☐

Quality of the CT Scan

Poor ☐Fair ☐Excellent ☐**CT #12**Type I ☐Type II A ☐Type II B ☐Type II C ☐Type III AB ☐Type III AC ☐Type III BC ☐Type IV ☐

Quality of the CT Scan

Poor ☐Fair ☐Excellent ☐**CT #13**Type I ☐Type II A ☐Type II B ☐Type II C ☐Type III AB ☐Type III AC ☐Type III BC ☐Type IV ☐

Quality of the CT Scan

Poor ☐Fair ☐Excellent ☐

CT #14Type I ☐Type II A ☐Type II B ☐Type II C ☐Type III AB ☐Type III AC ☐Type III BC ☐Type IV ☐

Quality of the CT Scan

Poor ☐Fair ☐Excellent ☐**CT #15**Type I ☐Type II A ☐Type II B ☐Type II C ☐Type III AB ☐Type III AC ☐Type III BC ☐Type IV ☐

Quality of the CT Scan

Poor ☐Fair ☐Excellent ☐**CT #16**Type I ☐Type II A ☐Type II B ☐Type II C ☐Type III AB ☐Type III AC ☐Type III BC ☐Type IV ☐

Quality of the CT Scan

Poor ☐Fair ☐Excellent ☐**CT #17**Type I ☐Type II A ☐Type II B ☐Type II C ☐Type III AB ☐Type III AC ☐Type III BC ☐Type IV ☐

Quality of the CT Scan

Poor ☐Fair ☐Excellent ☐

CT #18Type I ☐Type II A ☐Type II B ☐Type II C ☐Type III AB ☐Type III AC ☐Type III BC ☐Type IV ☐

Quality of the CT Scan

Poor ☐Fair ☐Excellent ☐**CT #19**Type I ☐Type II A ☐Type II B ☐Type II C ☐Type III AB ☐Type III AC ☐Type III BC ☐Type IV ☐

Quality of the CT Scan

Poor ☐Fair ☐Excellent ☐**CT #20**Type I ☐Type II A ☐Type II B ☐Type II C ☐Type III AB ☐Type III AC ☐Type III BC ☐Type IV ☐

Quality of the CT Scan

Poor ☐Fair ☐Excellent ☐**CT #21**Type I ☐Type II A ☐Type II B ☐Type II C ☐Type III AB ☐Type III AC ☐Type III BC ☐Type IV ☐

Quality of the CT Scan

Poor ☐Fair ☐Excellent ☐**CT #22**Type I ☐Type II A ☐Type II B ☐Type II C ☐Type III AB ☐Type III AC ☐Type III BC ☐Type IV ☐

Quality of the CT Scan

Poor ☐Fair ☐Excellent ☐

CT #23Type I ☐Type II A ☐Type II B ☐Type II C ☐Type III AB ☐Type III AC ☐Type III BC ☐Type IV ☐

Quality of the CT Scan

Poor ☐Fair ☐Excellent ☐**CT #24**Type I ☐Type II A ☐Type II B ☐Type II C ☐Type III AB ☐Type III AC ☐Type III BC ☐Type IV ☐

Quality of the CT Scan

Poor ☐Fair ☐Excellent ☐**CT #25**Type I ☐Type II A ☐Type II B ☐Type II C ☐Type III AB ☐Type III AC ☐Type III BC ☐Type IV ☐

Quality of the CT Scan

Poor ☐Fair ☐Excellent ☐**CT #26**Type I ☐Type II A ☐Type II B ☐Type II C ☐Type III AB ☐Type III AC ☐Type III BC ☐Type IV ☐

Quality of the CT Scan

Poor ☐Fair ☐Excellent ☐**CT #27**Type I ☐Type II A ☐Type II B ☐Type II C ☐Type III AB ☐Type III AC ☐Type III BC ☐Type IV ☐

Quality of the CT Scan

Poor ☐Fair ☐Excellent ☐

CT #28Type I ☐Type II A ☐Type II B ☐Type II C ☐Type III AB ☐Type III AC ☐Type III BC ☐Type IV ☐

Quality of the CT Scan

Poor ☐Fair ☐Excellent ☐**CT #29**Type I ☐Type II A ☐Type II B ☐Type II C ☐Type III AB ☐Type III AC ☐Type III BC ☐Type IV ☐

Quality of the CT Scan

Poor ☐Fair ☐Excellent ☐**CT #30**Type I ☐Type II A ☐Type II B ☐Type II C ☐Type III AB ☐Type III AC ☐Type III BC ☐Type IV ☐

Quality of the CT Scan

Poor ☐Fair ☐Excellent ☐

How familiar are you with Sanders classification?

Not at all ☐

Moderately ☐

Very ☐

Did you experience any difficulties with the quality of the radiographs, if so please indicate which ones.

Did you experience any problems with the questionnaire?

Comments

Thank You For Your Time

Appendix B

Appendix C

**Memorial University Of Newfoundland
Faculty of Medicine
Research Project Questionnaire**

The questionnaire below will take approximately 20-30 minutes to complete. Could you please examine each of the radiographs of subtrochanteric fractures provided, and check box, corresponding to Russel-Taylor class in which you believe the fracture pattern fits. Thank you in advance for your time.

1. Radiograph # 1

Type IA ☐ Type IIA ☐ Type IB ☐ Type IIB ☐
Quality of the Radiograph Poor ☐ Fair ☐ Excellent ☐

2. Radiograph # 2

Type IA ☐ Type IIA ☐ Type IB ☐ Type IIB ☐
Quality of the Radiograph Poor ☐ Fair ☐ Excellent ☐

3. Radiograph # 3

Type IA ☐ Type IIA ☐ Type IB ☐ Type IIB ☐
Quality of the Radiograph Poor ☐ Fair ☐ Excellent ☐

4. Radiograph # 4

Type IA ☐ Type IIA ☐ Type IB ☐ Type IIB ☐
Quality of the Radiograph Poor ☐ Fair ☐ Excellent ☐

5. Radiograph # 5

Type IA ☐ Type IIA ☐ Type IB ☐ Type IIB ☐
Quality of the Radiograph Poor ☐ Fair ☐ Excellent ☐

6. Radiograph # 6

Type IA ☐

Type IIA ☐

Type IB ☐

Type IIB ☐

Quality of the Radiograph Poor ☐

Fair ☐

Excellent ☐

7. Radiograph # 7

Type IA ☐

Type IIA ☐

Type IB ☐

Type IIB ☐

Quality of the Radiograph Poor ☐

Fair ☐

Excellent ☐

8. Radiograph # 8

Type IA ☐

Type IIA ☐

Type IB ☐

Type IIB ☐

Quality of the Radiograph Poor ☐

Fair ☐

Excellent ☐

9. Radiograph # 9

Type IA ☐

Type IIA ☐

Type IB ☐

Type IIB ☐

Quality of the Radiograph Poor ☐

Fair ☐

Excellent ☐

10. Radiograph # 10

Type IA ☐

Type IIA ☐

Type IB ☐

Type IIB ☐

Quality of the Radiograph Poor ☐

Fair ☐

Excellent ☐

11. Radiograph # 11

Type IA ☐

Type IIA ☐

Type IB ☐

Type IIB ☐

Quality of the Radiograph Poor ☐

Fair ☐

Excellent ☐

12. Radiograph # 12

Type IA ☐

Type IIA ☐

Type IB ☐

Type IIB ☐

Quality of the Radiograph Poor ☐

Fair ☐

Excellent ☐

13. Radiograph # 13

Type IA ☐

Type IIA ☐

Type IB ☐

Type IIB ☐

Quality of the Radiograph Poor ☐

Fair ☐

Excellent ☐

14. Radiograph # 14

Type IA ☐

Type IIA ☐

Type IB ☐

Type IIB ☐

Quality of the Radiograph Poor ☐

Fair ☐

Excellent ☐

15. Radiograph # 15

Type IA ☐

Type IIA ☐

Type IB ☐

Type IIB ☐

Quality of the Radiograph Poor ☐

Fair ☐

Excellent ☐

16. Radiograph # 16

Type IA ☐

Type IIA ☐

Type IB ☐

Type IIB ☐

Quality of the Radiograph Poor ☐

Fair ☐

Excellent ☐

How familiar are you with Russel-Taylor classification?

Not at all ☐

Moderately ☐

Very ☐

Did you experience any difficulties with the quality of the radiographs, if so please indicate which ones.

Did you experience any problems with the questionnaire?

Comments

Thank You For Your Time

Appendix D

Appendix E

**Memorial University Of Newfoundland
Faculty of Medicine
Research Project Questionnaire**

*The questionnaire below will take approximately 20-30 minutes to complete. Could you please examine each of the CT Scans and using the ruler provided measure the intrapedicular(IP) distance and anterior-posterior(AP) distance of the canal from the image marked at levels L3-4, L4-5 and L5-S1 and classify the stenosis accordingly.
Thank you in advance for your time.*

FILM # 1

L3-4 AP distance _____mm
IP distance _____mm

L4-5 AP distance _____mm
IP distance _____mm

L5-S1 AP distance _____mm
IP distance _____mm

Normal Canal ☐

Mild Stenosis ☐

Moderate Stenosis ☐

Severe Stenosis ☐

Quality of the CT Scan Poor ☐

Fair ☐

Excellent ☐

FILM # 2

L3-4 AP distance _____mm
IP distance _____mm

L4-5 AP distance _____mm
IP distance _____mm

L5-S1 AP distance _____mm
IP distance _____mm

Normal Canal ☐

Mild Stenosis ☐

Moderate Stenosis ☐

Severe Stenosis ☐

Quality of the CT Scan Poor ☐

Fair ☐

Excellent ☐

FILM # 3

L3-4 AP distance _____mm
IP distance _____mm

L4-5 AP distance _____mm
IP distance _____mm

L5-S1 AP distance _____mm
IP distance _____mm

Normal Canal ☐
Moderate Stenosis ☐

Mild Stenosis ☐
Severe Stenosis ☐

Quality of the CT Scan Poor ☐

Fair ☐

Excellent ☐

FILM # 4

L3-4 AP distance _____mm
IP distance _____mm

L4-5 AP distance _____mm
IP distance _____mm

L5-S1 AP distance _____mm
IP distance _____mm

Normal Canal ☐
Moderate Stenosis ☐

Mild Stenosis ☐
Severe Stenosis ☐

Quality of the CT Scan Poor ☐

Fair ☐

Excellent ☐

FILM # 5

L3-4 AP distance _____mm
IP distance _____mm

L4-5 AP distance _____mm
IP distance _____mm

L5-S1 AP distance _____mm
IP distance _____mm

Normal Canal ☐
Moderate Stenosis ☐

Mild Stenosis ☐
Severe Stenosis ☐

Quality of the CT Scan Poor ☐

Fair ☐

Excellent ☐

FILM #6

L3-4 AP distance _____mm
IP distance _____mm

L4-5 AP distance _____mm
IP distance _____mm

L5-S1 AP distance _____mm
IP distance _____mm

Normal Canal ☐
Moderate Stenosis ☐

Mild Stenosis ☐
Severe Stenosis ☐

Quality of the CT Scan Poor ☐ Fair ☐ Excellent ☐

FILM #7

L3-4 AP distance _____mm
IP distance _____mm

L4-5 AP distance _____mm
IP distance _____mm

L5-S1 AP distance _____mm
IP distance _____mm

Normal Canal ☐
Moderate Stenosis ☐

Mild Stenosis ☐
Severe Stenosis ☐

Quality of the CT Scan Poor ☐ Fair ☐ Excellent ☐

FILM #8

L3-4 AP distance _____mm
IP distance _____mm

L4-5 AP distance _____mm
IP distance _____mm

L5-S1 AP distance _____mm
IP distance _____mm

Normal Canal ☐
Moderate Stenosis ☐

Mild Stenosis ☐
Severe Stenosis ☐

Quality of the CT Scan MRI Poor ☐ Fair ☐
Excellent ☐

FILM #9

L3-4 AP distance _____mm
IP distance _____mm

L4-5 AP distance _____mm
IP distance _____mm

L5-S1 AP distance _____mm
IP distance _____mm

Normal Canal ☐
Moderate Stenosis ☐

Mild Stenosis ☐
Severe Stenosis ☐

Quality of the CT Scan or MRI Poor ☐ Fair ☐
Excellent ☐

FILM # 10

L3-4 AP distance _____mm
IP distance _____mm

L4-5 AP distance _____mm
IP distance _____mm

L5-S1 AP distance _____mm
IP distance _____mm

Normal Canal ☐
Moderate Stenosis ☐

Mild Stenosis ☐
Severe Stenosis ☐

Quality of the CT Scan or MRI Poor ☐ Fair ☐
Excellent ☐

FILM # 11

L3-4 AP distance _____mm
IP distance _____mm

L4-5 AP distance _____mm
IP distance _____mm

L5-S1 AP distance _____mm
IP distance _____mm

Normal Canal ☐
Moderate Stenosis ☐

Mild Stenosis ☐
Severe Stenosis ☐

Quality of the CT Scan or MRI Poor ☐
Excellent ☐

Fair ☐

FILM # 12

L3-4 AP distance _____mm
IP distance _____mm

L4-5 AP distance _____mm
IP distance _____mm

L5-S1 AP distance _____mm
IP distance _____mm

Normal Canal ☐
Moderate Stenosis ☐

Mild Stenosis ☐
Severe Stenosis ☐

Quality of the CT Scan or MRI Poor ☐
Excellent ☐

Fair ☐

FILM #13

L3-4 AP distance _____mm
IP distance _____mm

L4-5 AP distance _____mm
IP distance _____mm

L5-S1 AP distance _____mm
IP distance _____mm

Normal Canal ☐
Moderate Stenosis ☐

Mild Stenosis ☐
Severe Stenosis ☐

Quality of the CT Scan or MRI Poor ☐
Excellent ☐

Fair ☐

FILM #14

L3-4 AP distance _____mm
IP distance _____mm

L4-5 AP distance _____mm
IP distance _____mm

L5-S1 AP distance _____mm
IP distance _____mm

Normal Canal ☐
Moderate Stenosis ☐

Mild Stenosis ☐
Severe Stenosis ☐

Quality of the CT Scan or MRI Poor ☐ Fair ☐
Excellent ☐

FILM #15

L3-4 AP distance _____mm
IP distance _____mm

L4-5 AP distance _____mm
IP distance _____mm

L5-S1 AP distance _____mm
IP distance _____mm

Normal Canal ☐
Moderate Stenosis ☐

Mild Stenosis ☐
Severe Stenosis ☐

Quality of the CT Scan or MRI Poor ☐ Fair ☐
Excellent ☐

FILM #16

L3-4 AP distance _____mm
IP distance _____mm

L4-5 AP distance _____mm
IP distance _____mm

L5-S1 AP distance _____mm
IP distance _____mm

Normal Canal ☐
Moderate Stenosis ☐

Mild Stenosis ☐
Severe Stenosis ☐

Quality of the CT Scan or MRI Poor ☐ Fair ☐
Excellent ☐

FILM # 17

L3-4 AP distance _____mm
IP distance _____mm

L4-5 AP distance _____mm
IP distance _____mm

L5-S1 AP distance _____mm
IP distance _____mm

Normal Canal ☐
Moderate Stenosis ☐

Mild Stenosis ☐
Severe Stenosis ☐

Quality of the CT Scan or MRI Poor ☐ Fair ☐
Excellent ☐

FILM # 18

L3-4 AP distance _____mm
IP distance _____mm

L4-5 AP distance _____mm
IP distance _____mm

L5-S1 AP distance _____mm
IP distance _____mm

Normal Canal ☐
Moderate Stenosis ☐

Mild Stenosis ☐
Severe Stenosis ☐

Quality of the CT Scan or MRI Poor ☐ Fair ☐
Excellent ☐

FILM # 19

L3-4 AP distance _____mm
IP distance _____mm

L4-5 AP distance _____mm
IP distance _____mm

L5-S1 AP distance _____mm
IP distance _____mm

Normal Canal ☐
Moderate Stenosis ☐

Mild Stenosis ☐
Severe Stenosis ☐

Quality of the CT Scan or MRI Poor ☐ Fair ☐
Excellent ☐

FILM #20

L3-4 AP distance _____mm
IP distance _____mm

L4-5 AP distance _____mm
IP distance _____mm

L5-S1 AP distance _____mm
IP distance _____mm

Normal Canal ☐
Moderate Stenosis ☐

Mild Stenosis ☐
Severe Stenosis ☐

Quality of the CT Scan or MRI Poor ☐
Excellent ☐

Fair ☐

FILM #21

L3-4 AP distance _____mm
IP distance _____mm

L4-5 AP distance _____mm
IP distance _____mm

L5-S1 AP distance _____mm
IP distance _____mm

Normal Canal ☐
Moderate Stenosis ☐

Mild Stenosis ☐
Severe Stenosis ☐

Quality of the CT Scan or MRI Poor ☐
Excellent ☐

Fair ☐

FILM #22

L3-4 AP distance _____mm
IP distance _____mm

L4-5 AP distance _____mm
IP distance _____mm

L5-S1 AP distance _____mm
IP distance _____mm

Normal Canal ☐
Moderate Stenosis ☐

Mild Stenosis ☐
Severe Stenosis ☐

Quality of the CT Scan Poor ☐

Fair ☐

Excellent ☐

FILM #23

L3-4 AP distance _____mm
IP distance _____mm

L4-5 AP distance _____mm
IP distance _____mm

L5-S1 AP distance _____mm
IP distance _____mm

Normal Canal ☐
Moderate Stenosis ☐

Mild Stenosis ☐
Severe Stenosis ☐

Quality of the CT Scan Poor ☐ Fair ☐ Excellent ☐

FILM # 24

L3-4 AP distance _____mm
IP distance _____mm

L4-5 AP distance _____mm
IP distance _____mm

L5-S1 AP distance _____mm
IP distance _____mm

Normal Canal ☐
Moderate Stenosis ☐

Mild Stenosis ☐
Severe Stenosis ☐

Quality of the CT Scan Poor ☐ Fair ☐ Excellent ☐

FILM # 25

L3-4 AP distance _____mm
IP distance _____mm

L4-5 AP distance _____mm
IP distance _____mm

L5-S1 AP distance _____mm
IP distance _____mm

Normal Canal ☐
Moderate Stenosis ☐

Mild Stenosis ☐
Severe Stenosis ☐

Quality of the CT Scan Poor ☐ Fair ☐ Excellent ☐

To which department do you belong?

Orthopedics ☐

Nuerosurgery ☐

Radiology ☐

How familiar are you with measurements of the spinal canal space?

Not at all ☐

Moderately ☐

Very ☐

Did you experience any difficulties with the quality of the radiographs, if so please indicate which ones.

Did you experience any problems with the questionnaire?

Comments

Thank You For Your Time

Appendix F

